PROJECT

Time Series Forecasting

Chetan R Deshpande

Content

Content	Page No. - Sparkling	Rose

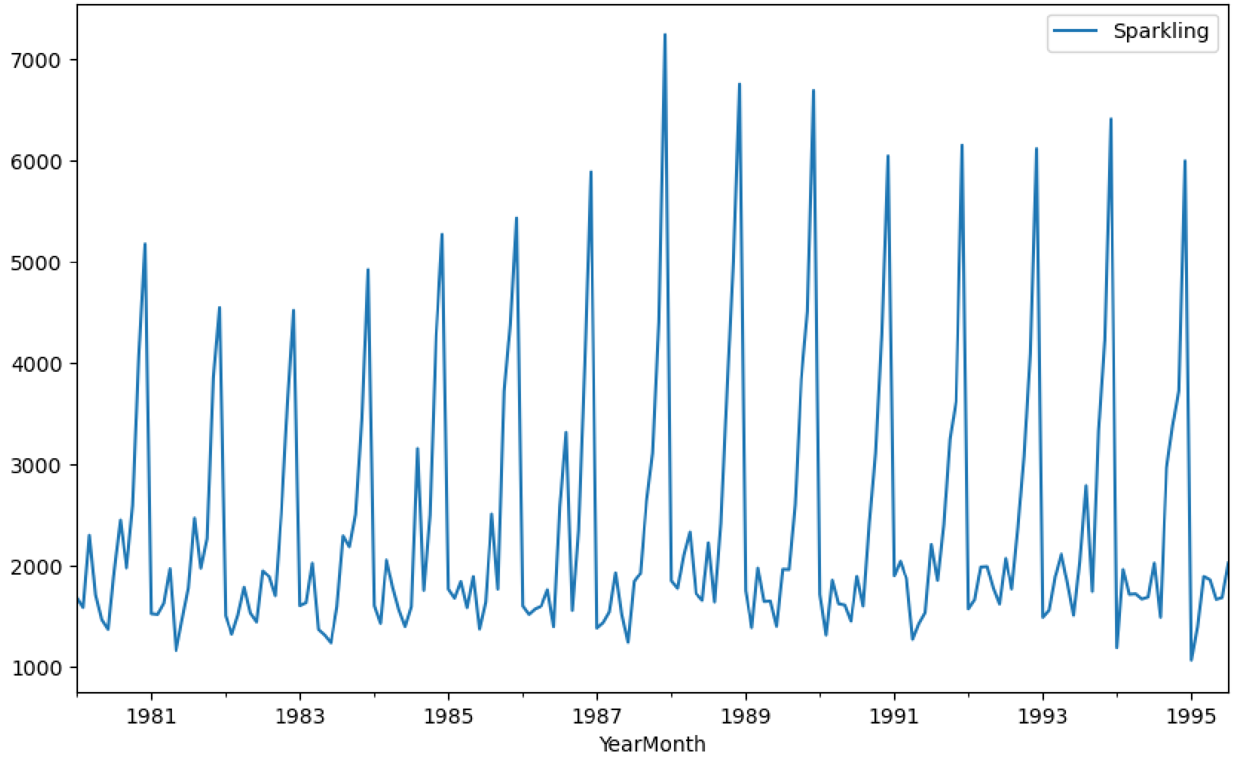
SPARKLING WINE

1. Read the data as an appropriate Time Series data and plot the data.

After reading the data using pd.read_csv, we will assign the YearMonth as the index. Further we have printed the head, which is the first 5 rows of the dataset to skim through the variable

	YearMonth	Sparkling
0	1980-01-01	1686
1	1980-02-01	1591
2	1980-03-01	2304
3	1980-04-01	1712
4	1980-05-01	1471

Now, we will plot the data and visually analyze the data of Sparkling Wine.



2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

- The shape of the data is (187, 1)
- The following is the information we got using .info()

  #   Column   Non-Null Count Dtype

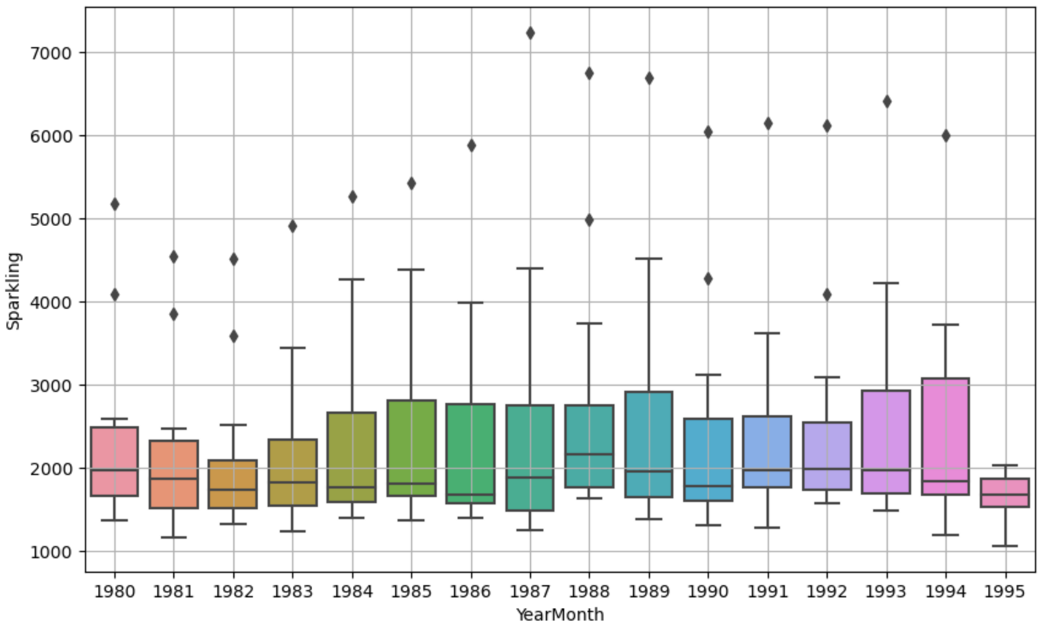
  --- ------   -------------- -----
   0   Sparkling 187 non-null   int64

  And we see that the variable is integer data type.
- Using .describe() we have printed the statistical summary of the data

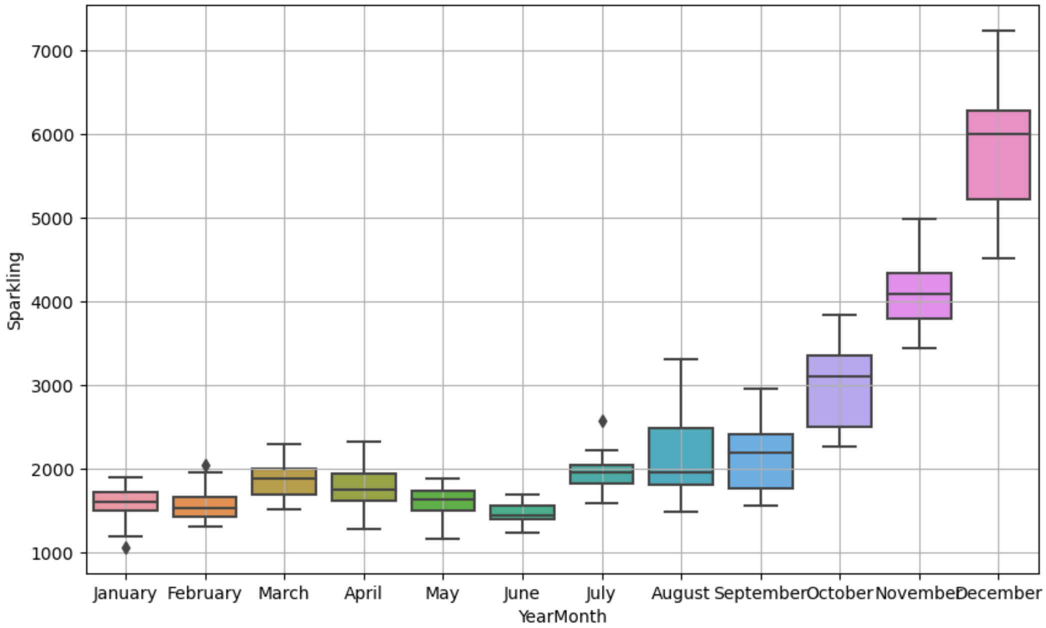
	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

- There is no null value present in the given dataset
  Sparkling   0
  dtype: int64

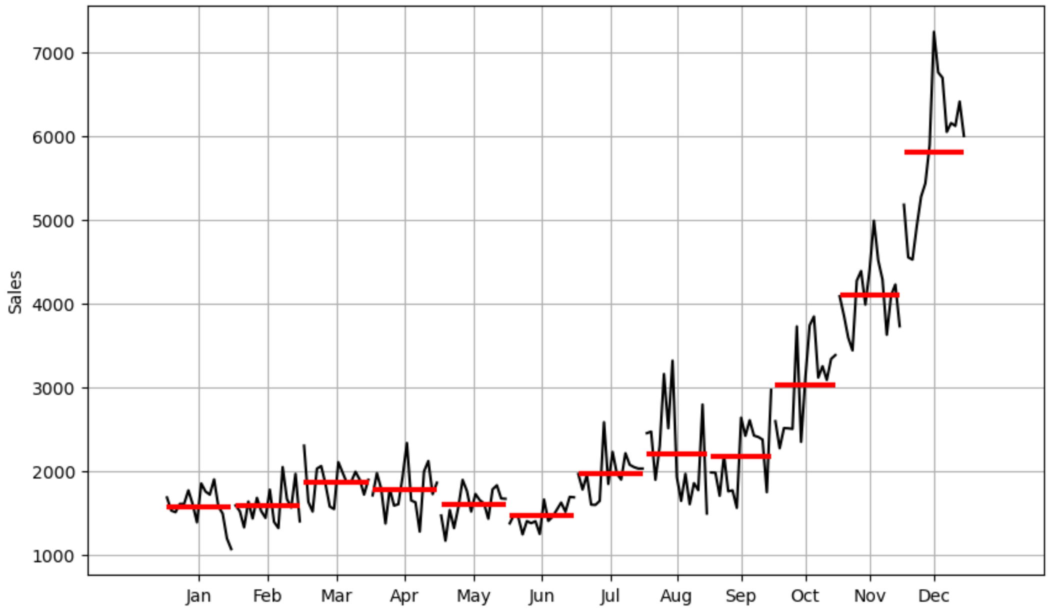
- The following is the yearly boxplot of the data



- The following is the monthly boxplot of the data



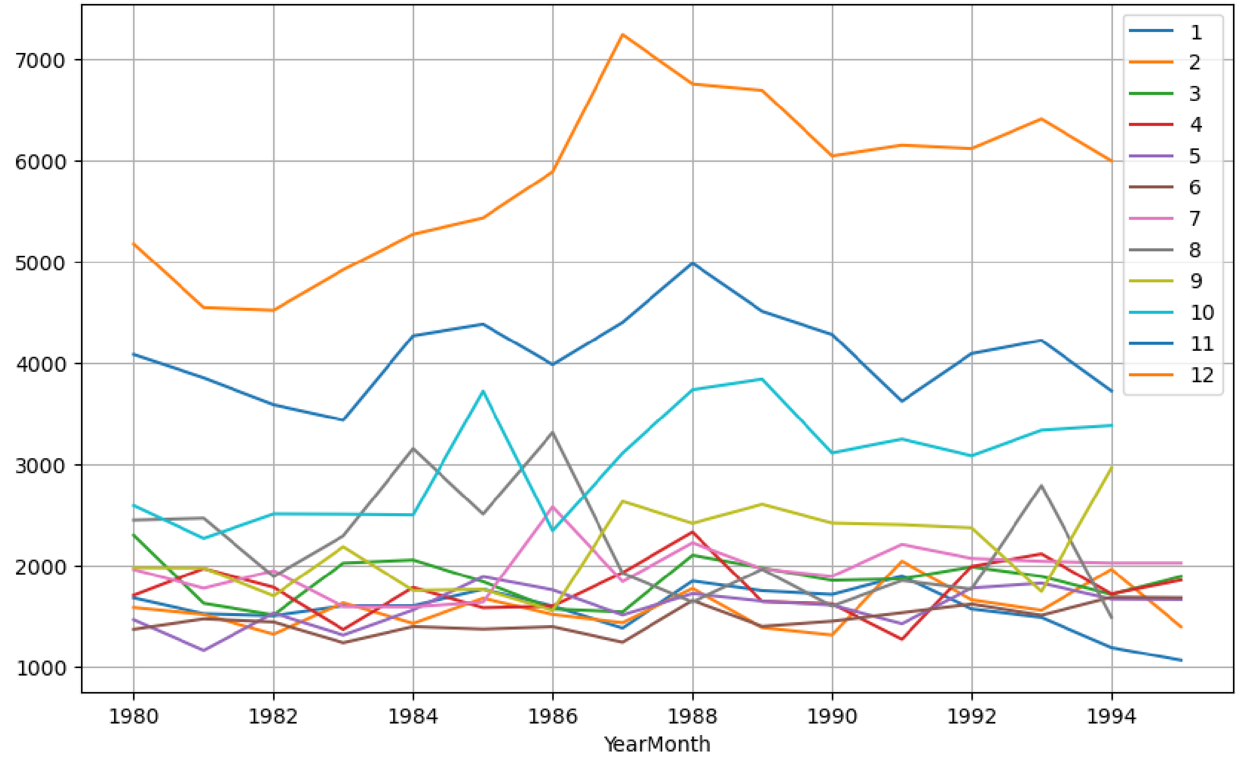
- Plotting a time series monthplot to understand the spread of accidents across different years and within different months across years



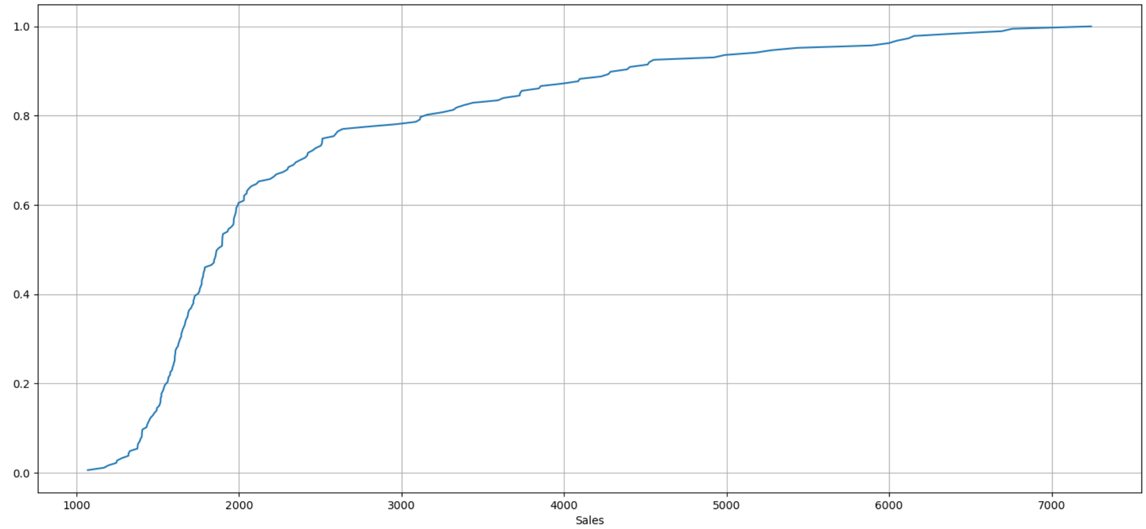
- To Plot a graph of monthly RetailSales across years, first we make a pivot table of the data

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

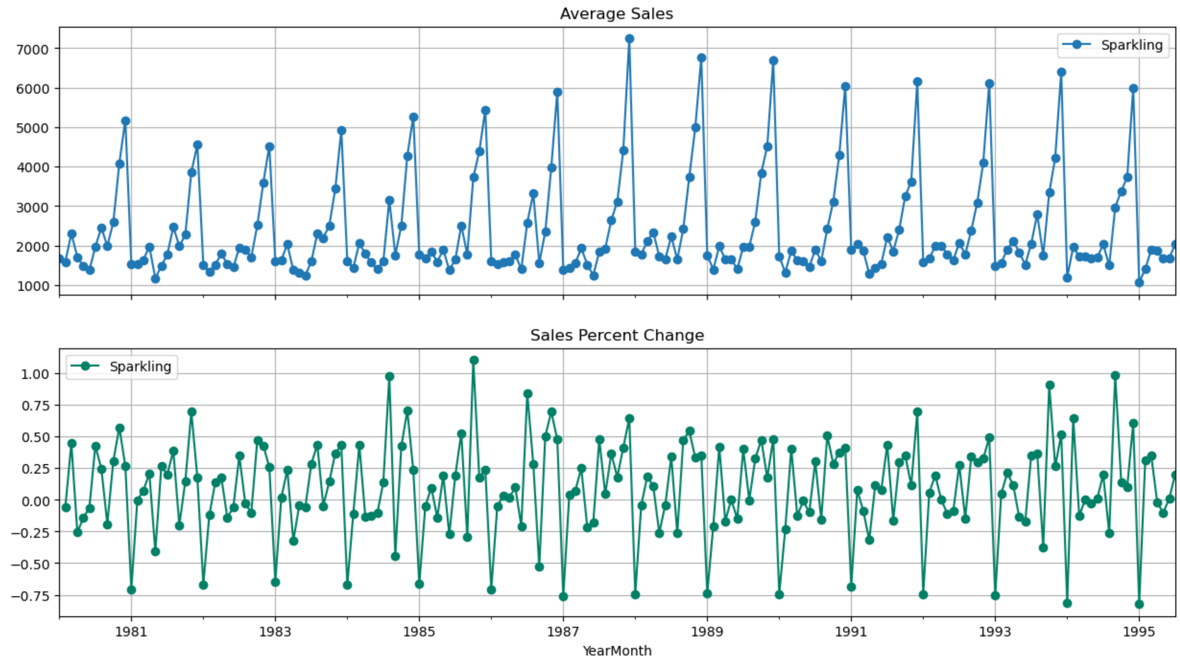
- Using the above table we have plotted a line graph to understand the data better.



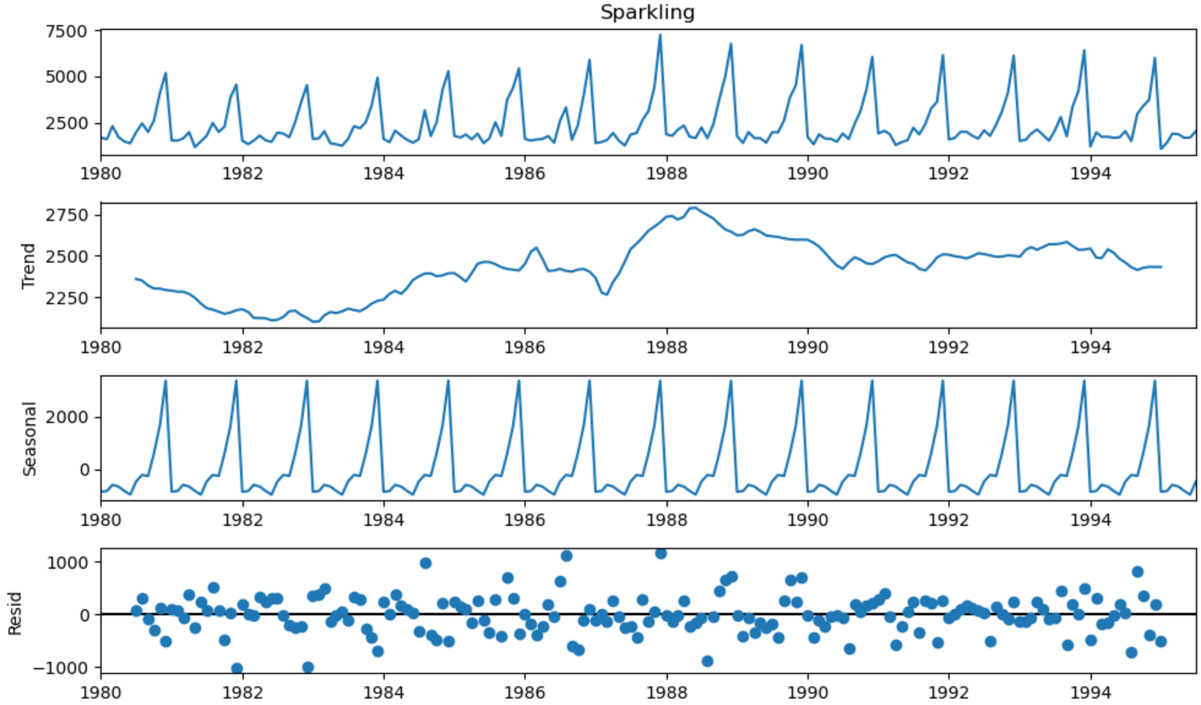
- Now we Empirical Cumulative Distribution



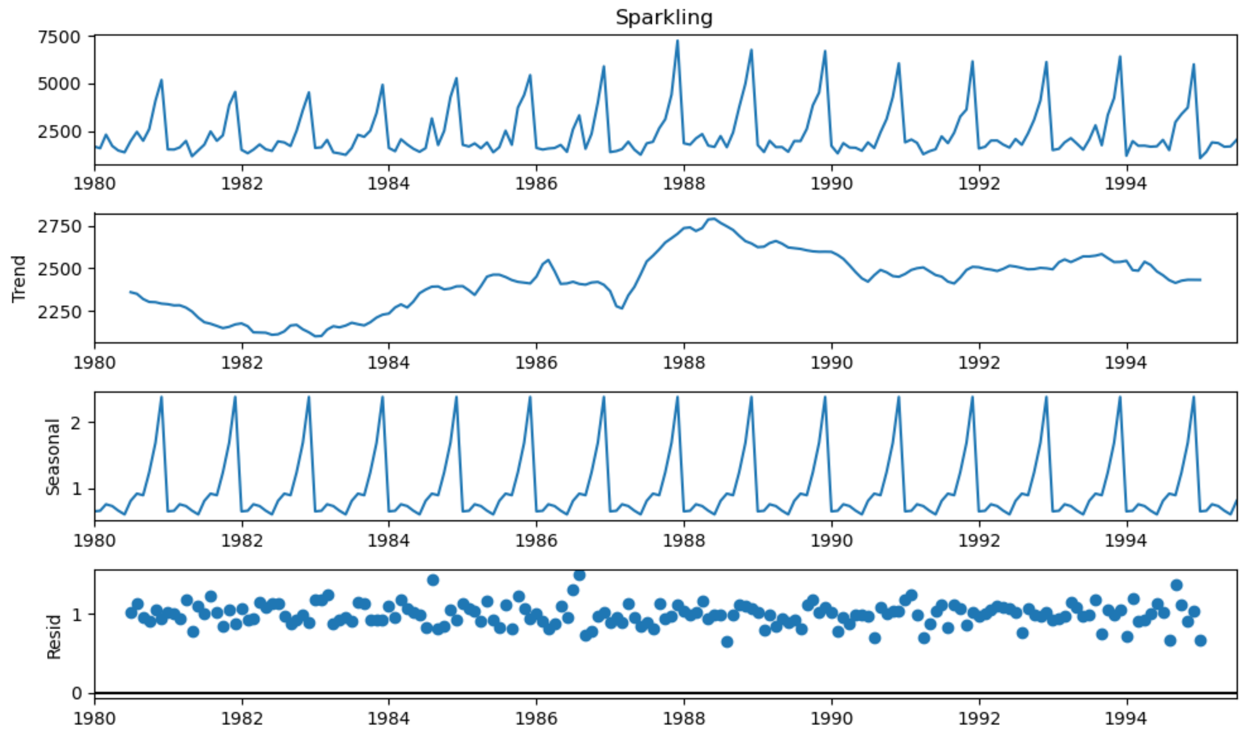
- Plotting the average RetailSales per month and the month on month percentage change of RetailSales



- In the next step, we use additive decomposition to decompose the Time Series and plot the different components



- Next, we use multiplicative decomposition to decompose the Time Series and plot the different components



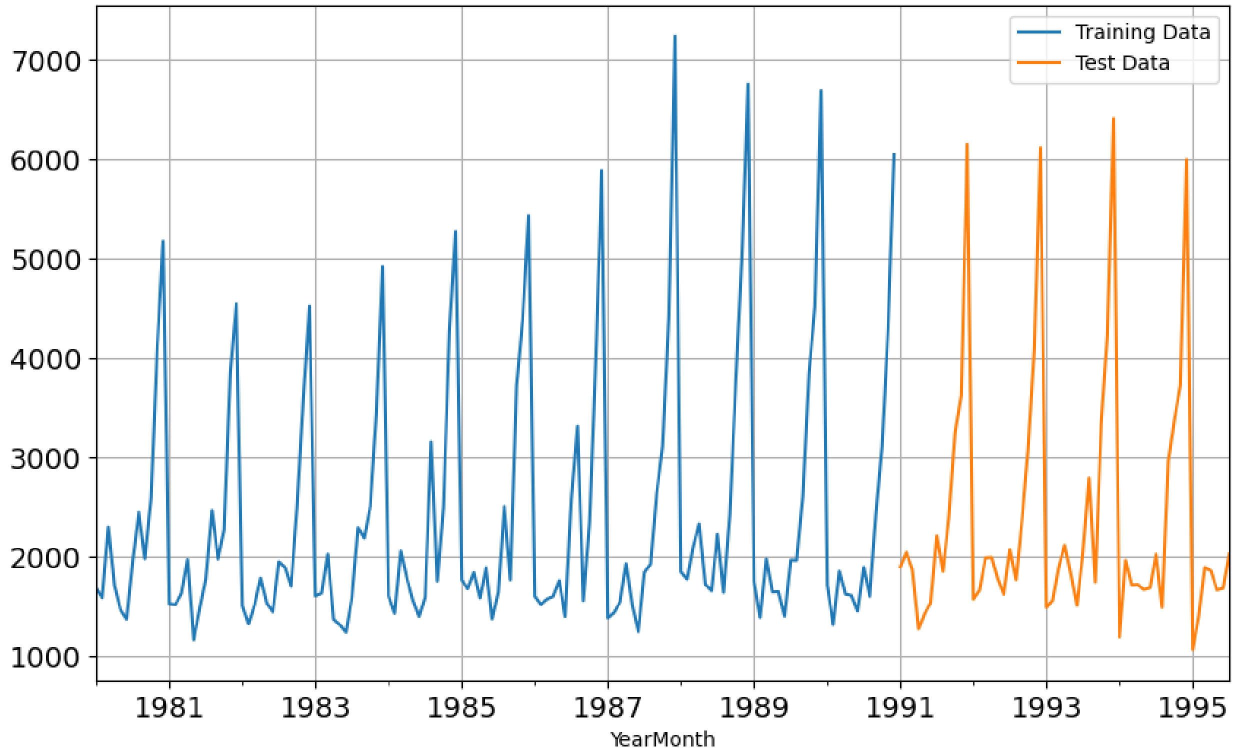
3. Split the data into training and testing. The test data should start in 1991.

Here, for training we use the data till 1990 and for testing the data post 1991 has been used.

The shape of the training data is - (132, 1)

The shape of the testing data is - (55, 1)

The following is the visualization for the train and test data split



4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

For this problem, we use multiple libraries like mean_squared_error from sklearn.metrics, ExponentialSmoothing, SimpleExpSmoothing, Holt from statsmodels.tsa.api and statsmodels.tools.eval_measures.

a. **Simple Exponential Smoothing**

The estimated parameters for SES is as follows

{'smoothing_level': 0.07028781460389553,

'smoothing_trend': nan,

'smoothing_seasonal': nan,

'damping_trend': nan,

'initial_level': 1763.9269926897714,

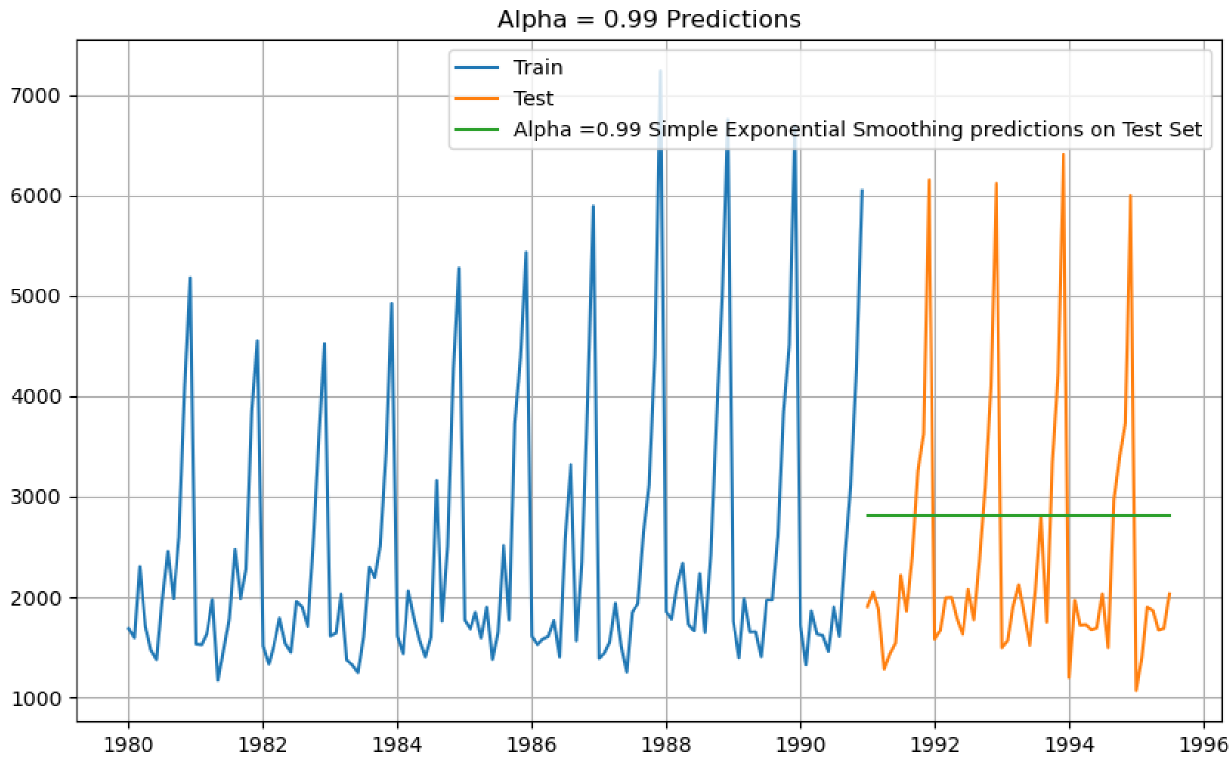
'initial_trend': nan,

'initial_seasons': array([], dtype=float64),

'use_boxcox': False,

'lamda': None,
'remove_bias': False}

Post this, we plot the following which is the simple exponential smoothing predictions on test data



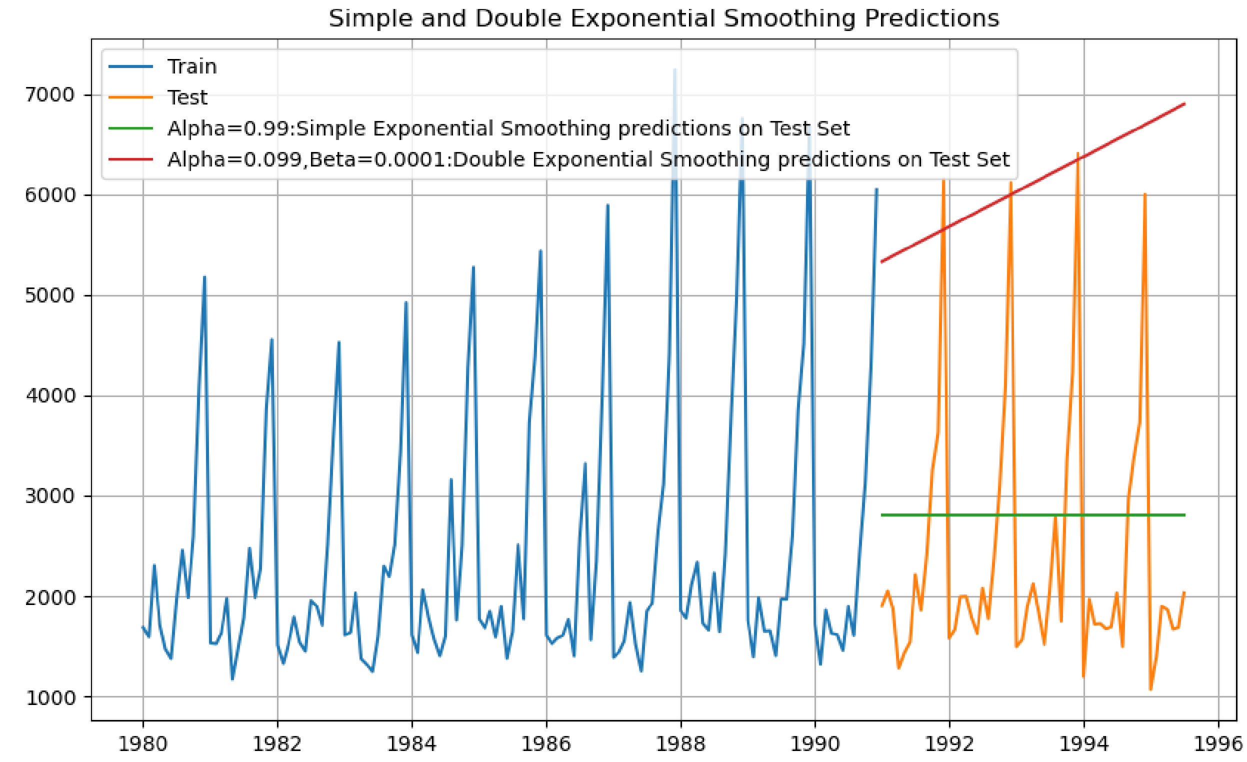
The RMSE for this model is 1338.0046232563645

b. **Double Exponential Smoothing**

The estimated parameters for DES is as follows

{'smoothing_level': 0.6638769379750992, 'smoothing_trend': 9.966252219085015e-05, 'smoothing_seasonal': nan, 'damping_trend': nan, 'initial_level': 1502.5681616946074, 'initial_trend': 29.018019327772173, 'initial_seasons': array([], dtype=float64), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}

Post this we will plot a graph to visualize the double exponential smoothing prediction on the test data



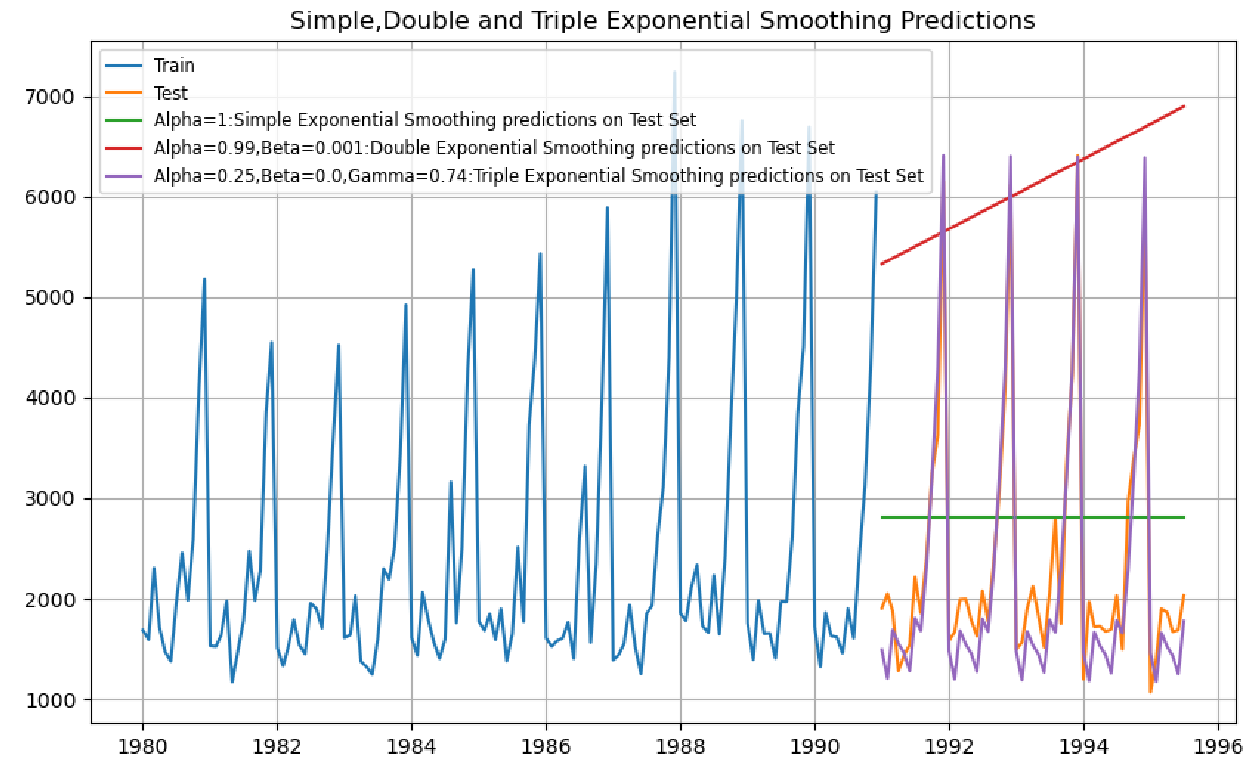
The RMSE that we have got for this prediction is 3949.9312976553306

c. Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors

These are the parameters obtained

{'smoothing_level': 0.1112722708441868, 'smoothing_trend': 0.012360804011805133, 'smoothing_seasonal': 0.4607176722812185, 'damping_trend': nan, 'initial_level': 2356.5780457745022, 'initial_trend': -0.10071063842436252, 'initial_seasons': array([-636.23320656, -722.98321108, -398.64408844, -473.43046147, -808.42474493, -815.34992566, -384.23070718, 72.99480164, -237.44227544, 272.32602348, 1541.37739846, 2590.07688887]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}

The following is the triple exponential smoothing prediction on the test data



The RMSE for this model is 378.9443254087722

Now, we will build different models and compare the accuracy metrics.

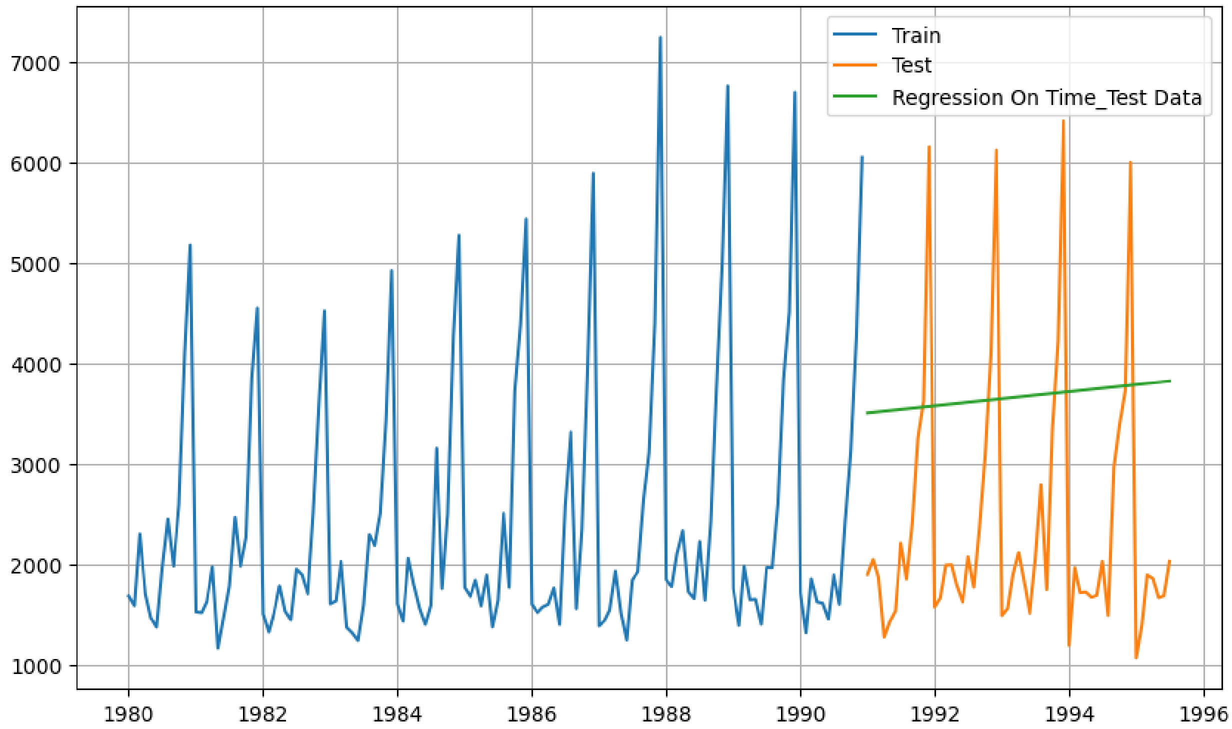
a. Linear Regression

Here, we see that we have successfully generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

	Sparkling	time
YearMonth		
1980-01-01	1686	1
1980-02-01	1591	2
1980-03-01	2304	3
1980-04-01	1712	4
1980-05-01	1471	5

Now that our training and test data has been modified, we go ahead and use LinearRegression to build the model on the training data and test the model on the test data.

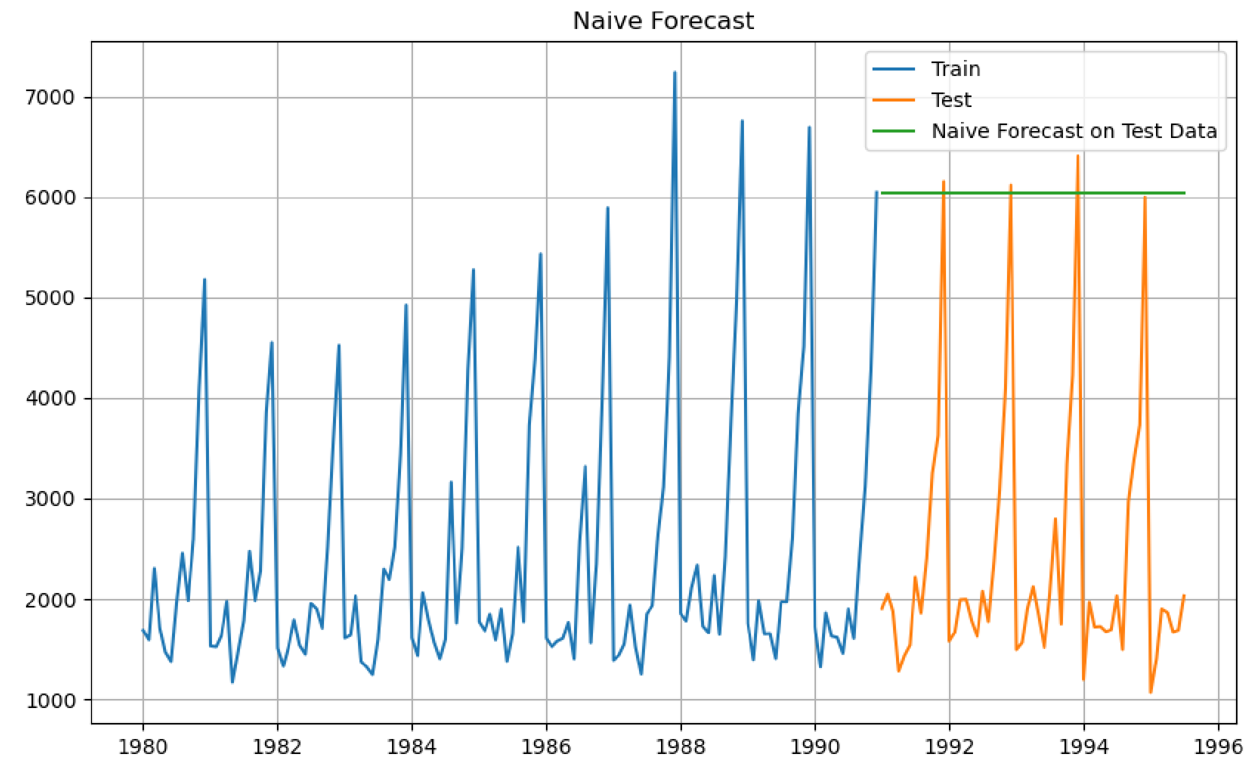
After we fit the data, the following is the output of the prediction



The RMSE for this model is RMSE is 1798.201

b. Naive Approach

After fitting the data, the following is the output

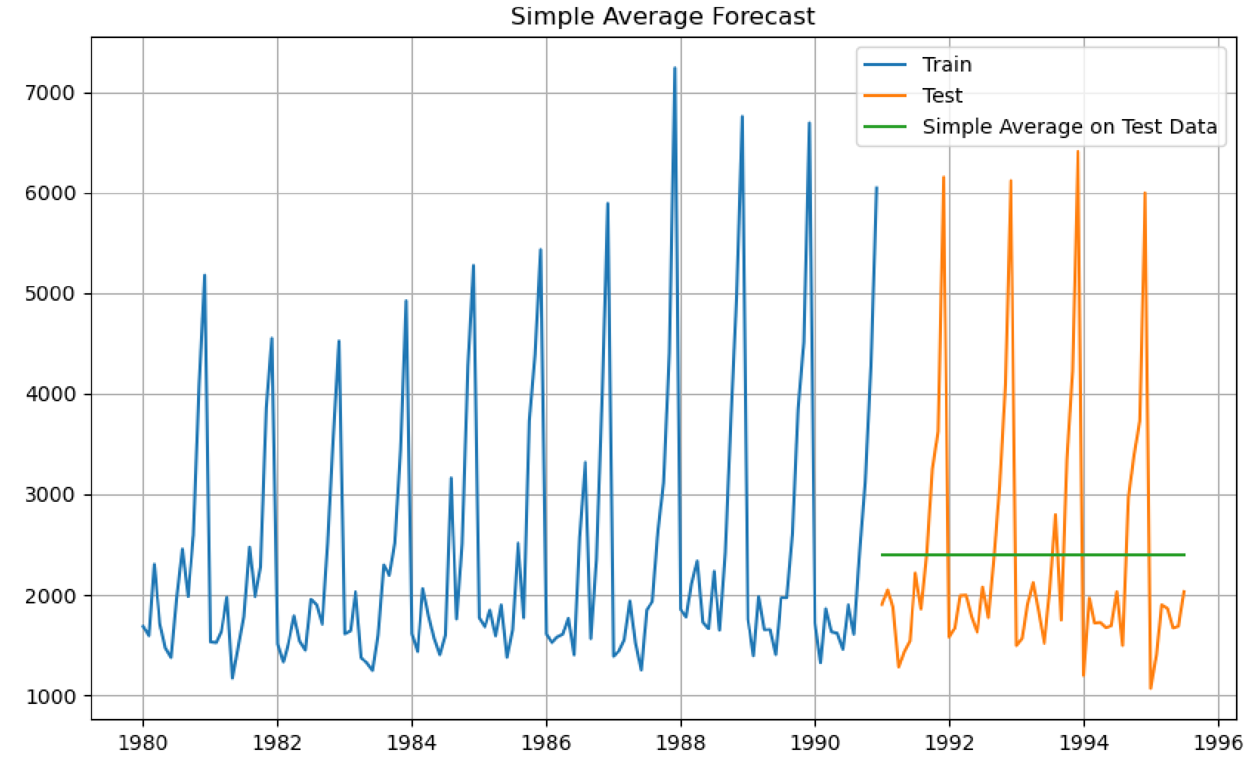


The RMSE for this approach is 3864.279

c. Simple Average

This model creates mean forecast to predict the future data

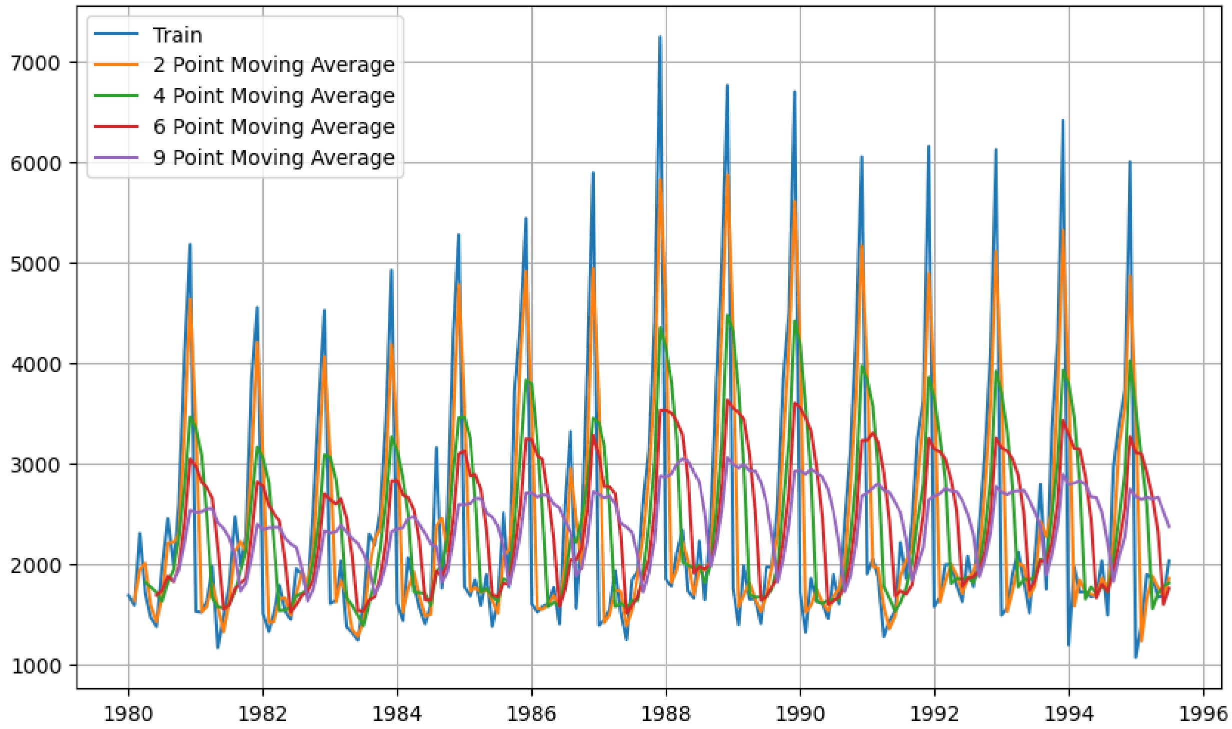
The following is the graph plotted to show the prediction on the test data



The RMSE value for this is 1275.082

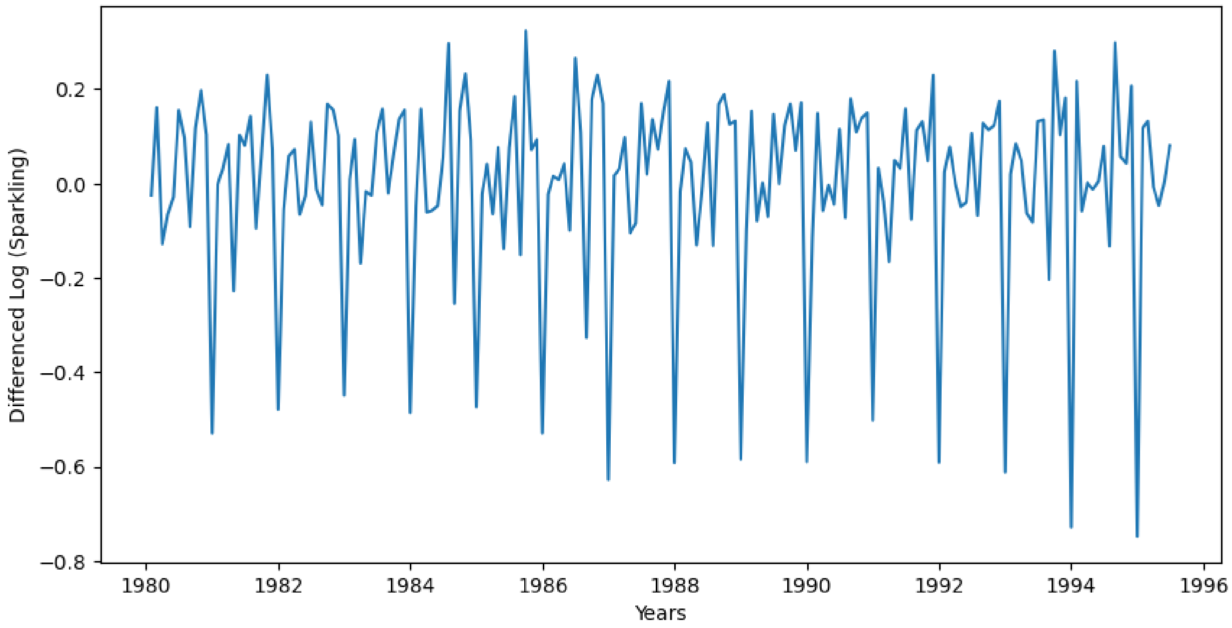
d. Moving Average(MA)

This is a statistical calculation that smooths out data points by averaging values within a sliding window, providing a trend or pattern representation while reducing noise.



Train
2 Point Moving Average
4 Point Moving Average
6 Point Moving Average
9 Point Moving Average
7000
6000
5000
4000
3000
2000
1000
1980
1982
1984
1986
1988
1990
1992
1994
1996

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.



We apply log to the data to get rid of the trend and seasonality.

6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

a. ARIMA

Building ARIMA model with best parameters p,d,q

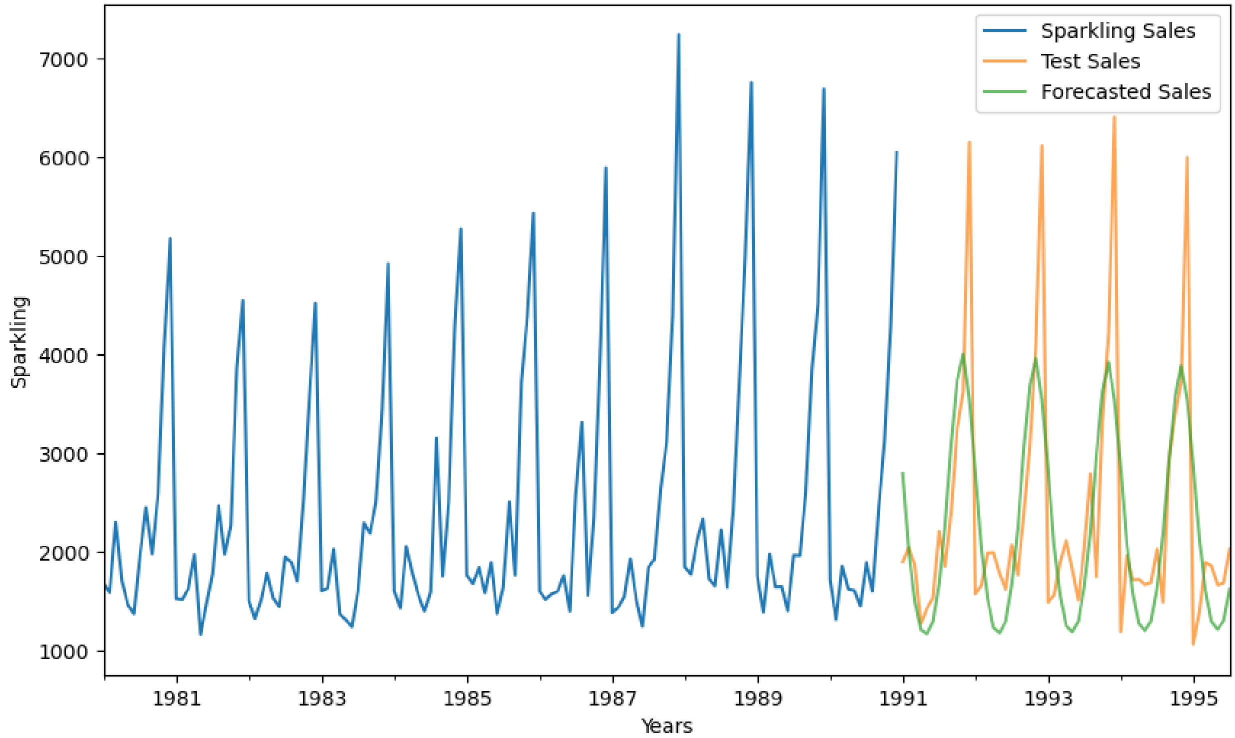
	param	AIC
13	(2, 0, 1)	-106.955371
17	(2, 1, 2)	-98.904295
12	(2, 0, 0)	-97.590175
2	(0, 0, 2)	-97.470702
7	(1, 0, 1)	-96.553751

The following is the best result summary

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(3, 0, 3)   Log Likelihood                  76.215
Date:                Sat, 30 Sep 2023   AIC                           -136.430
Time:                        17:45:50   BIC                           -113.367
Sample:                    01-01-1980   HQIC                          -127.058
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.3341      0.010    331.746      0.000       3.314       3.354
ar.L1          0.7373      0.383      1.927      0.054      -0.013       1.487
ar.L2          0.7244      0.660      1.098      0.272      -0.568       2.017
ar.L3         -0.9913      0.381     -2.603      0.009      -1.738      -0.245
ma.L1         -0.7928      0.372     -2.129      0.033      -1.523      -0.063
ma.L2         -0.7942      0.669     -1.187      0.235      -2.106       0.518
ma.L3          0.9838      0.396      2.481      0.013       0.207       1.761
sigma2         0.0173      0.003      5.661      0.000       0.011       0.023
==============================================================================
```

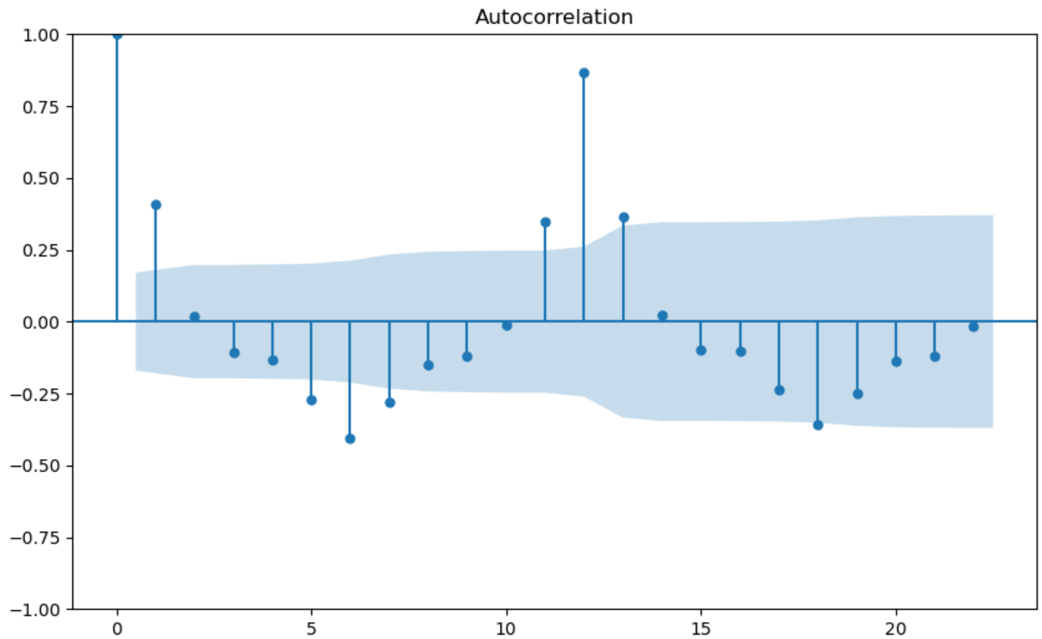
The RMSE for the above model is 940.865

The following is the plot for predicted sales



b. SARIMA Model

The start with plotting the ACF plot



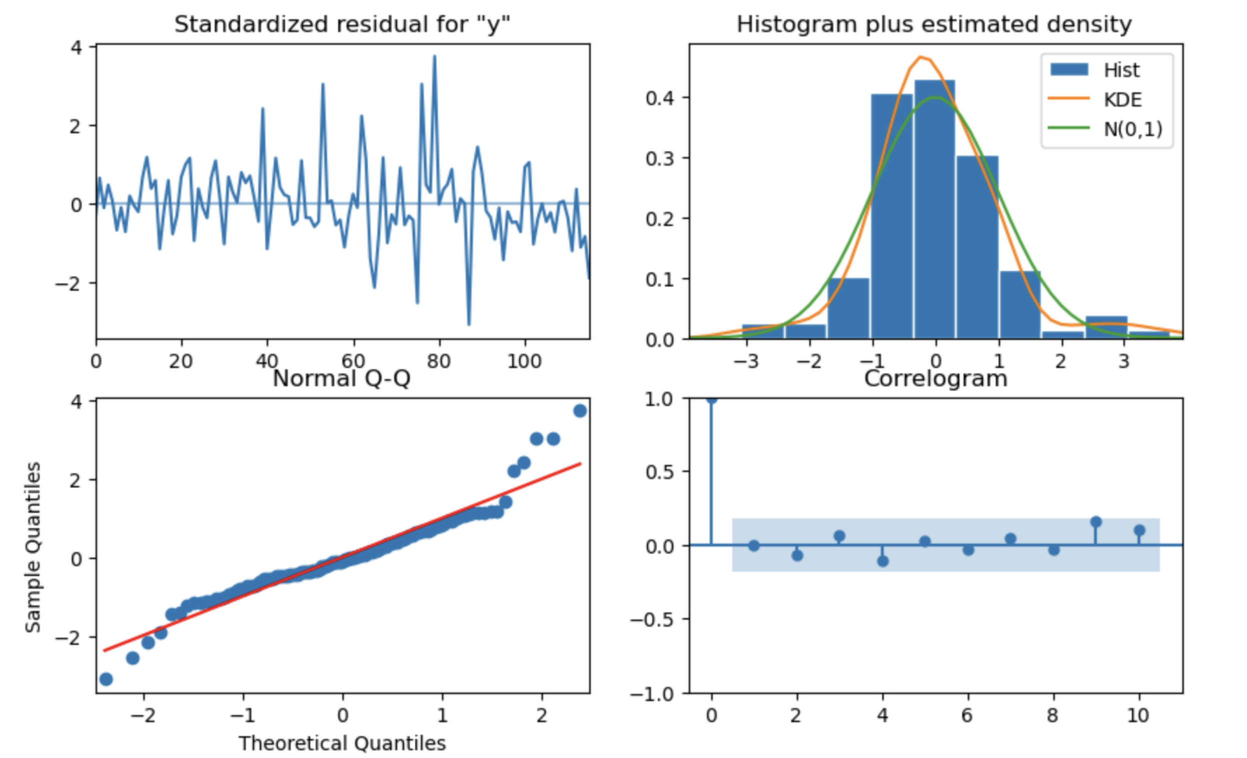
The following are some parameter combinations for Model...

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.678703
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888804
80	(2, 1, 2)	(2, 0, 2, 6)	1729.192580
17	(0, 1, 1)	(2, 0, 2, 6)	1741.647351
44	(1, 1, 1)	(2, 0, 2, 6)	1743.379780

The following is the summary for SARIMA

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(3, 0, 3)   Log Likelihood                  76.215
Date:                Sat, 30 Sep 2023   AIC                           -136.430
Time:                        17:45:50   BIC                           -113.367
Sample:                    01-01-1980   HQIC                          -127.058
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.3341      0.010    331.746      0.000       3.314       3.354
ar.L1          0.7373      0.383      1.927      0.054      -0.013       1.487
ar.L2          0.7244      0.660      1.098      0.272      -0.568       2.017
ar.L3         -0.9913      0.381     -2.603      0.009      -1.738      -0.245
ma.L1         -0.7928      0.372     -2.129      0.033      -1.523      -0.063
ma.L2         -0.7942      0.669     -1.187      0.235      -2.106       0.518
ma.L3          0.9838      0.396      2.481      0.013       0.207       1.761
sigma2         0.0173      0.003      5.661      0.000       0.011       0.023
==============================================================================
```

After which we have plotted the diagnostic plots



The RMSE value for this model is 601.2321647724793

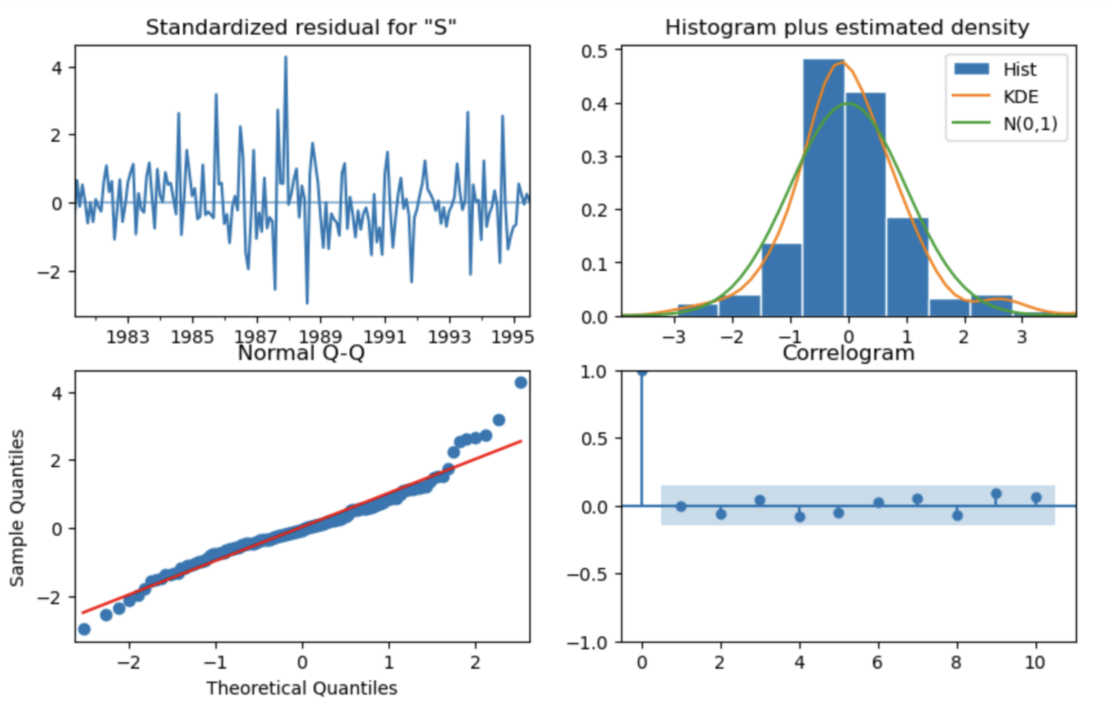
7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

	Test RMSE	RMSE
Alpha=0.99,SES	1338.004623	NaN
Alpha=1,Beta=0.0189:DES	3949.931298	NaN
Alpha=0.25,Beta=0.0,Gamma=0.74:TES	378.944325	NaN
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	403.125867	NaN
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	403.125867	NaN
NaiveModel	3864.279352	NaN
SimpleAverageModel	1275.081804	NaN
SimpleAverageModel	1275.081804	NaN
Best ARIMA Model : ARIMA(3,0,3)	NaN	940.864516
SARIMA(0,1,2)(2,0,2,6)	NaN	601.232165
SimpleAverageModel	1275.081804	NaN
SimpleAverageModel	1275.081804	NaN

The above is a table with all the models built along with their corresponding parameters and their respective values on the test data.

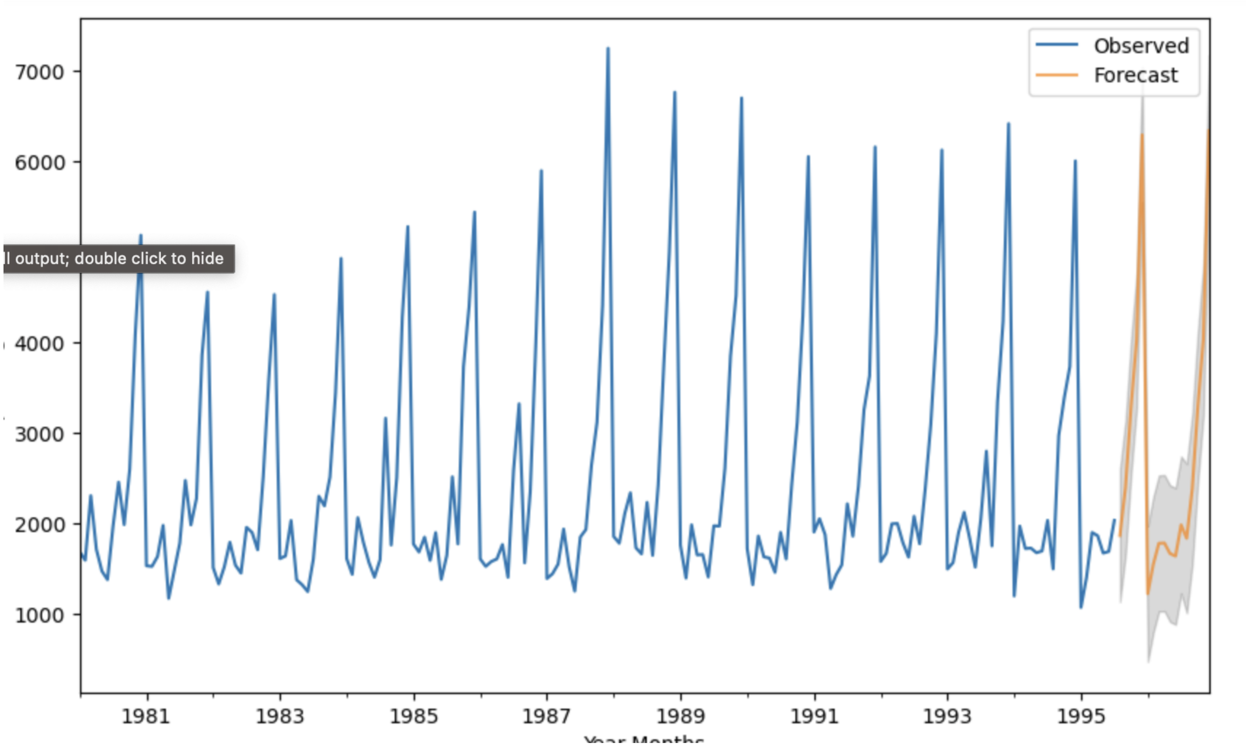
8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

```
CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                             Sparkling   No. Observations:              187
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood           -1258.196
Date:                           Sat, 30 Sep 2023   AIC                       2530.392
Time:                                   18:17:13   BIC                       2552.384
Sample:                               01-01-1980   HQIC                      2539.315
                                    - 07-01-1995
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8219      0.077    -10.657      0.000      -0.973      -0.671
ma.L2         -0.1100      0.079     -1.384      0.166      -0.266       0.046
ar.S.L6        0.0071      0.018      0.408      0.684      -0.027       0.041
ar.S.L12       1.0170      0.012     87.912      0.000       0.994       1.040
ma.S.L6       -0.0482      0.087     -0.556      0.578      -0.218       0.122
ma.S.L12      -0.6362      0.068     -9.324      0.000      -0.770      -0.502
sigma2      1.388e+05   1.09e+04     12.711      0.000    1.17e+05     1.6e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                56.44
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               1.24   Skew:                             0.62
Prob(H) (two-sided):                  0.42   Kurtosis:                         5.52
===================================================================================
```



The RMSE value is 531.9794862367713

The final plot of the prediction is on the entire data is



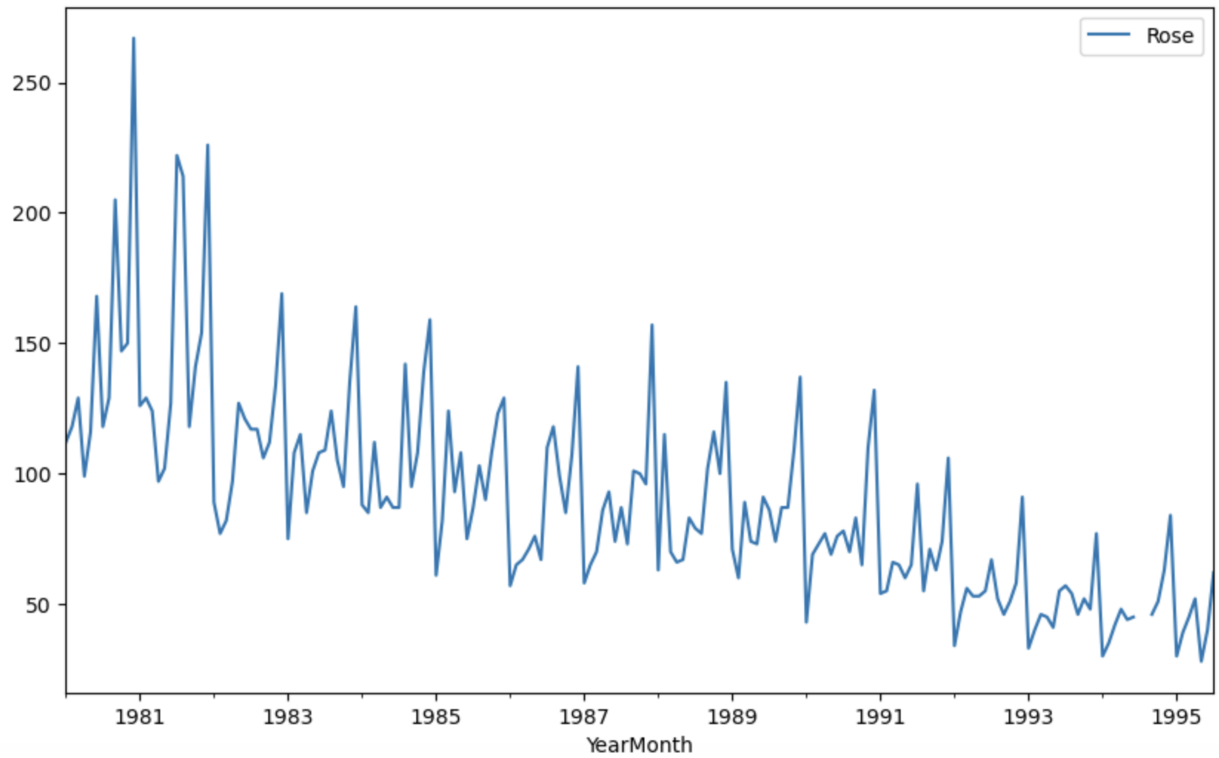
ROSE WINE

1. Read the data as an appropriate Time Series data and plot the data.

After reading the data using pd.read_csv, we will assign the YearMonth as the index. Further we have printed the head, which is the first 5 rows of the dataset to skim through the variable

	Rose
YearMonth	
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Now, we will plot the data and visually analyze the data of Rose Wine.



2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

- The shape of the data is (187, 1)
-
- There are 185b entries and 2 entries are found to be null.

 #   Column Non-Null Count Dtype

--- ------ -------------- -----

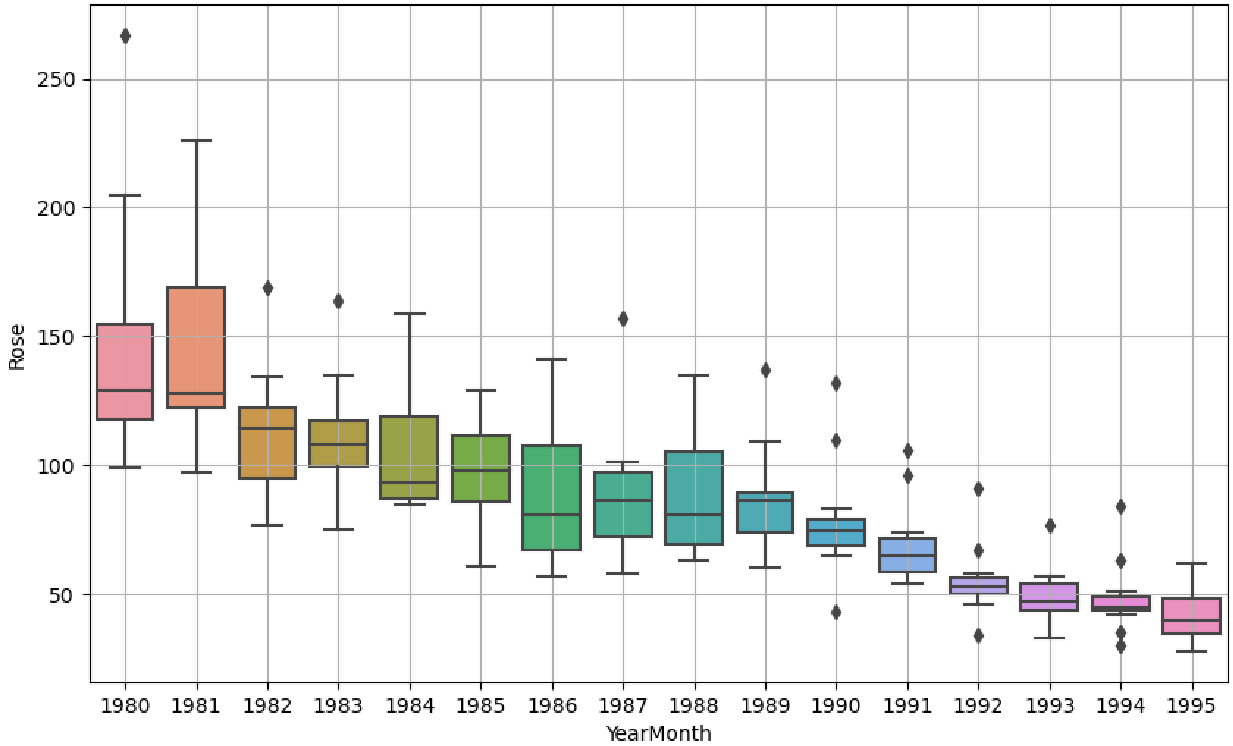
 0   Rose   185 non-null   float64

dtypes: float64(1)
memory usage: 2.9 KB

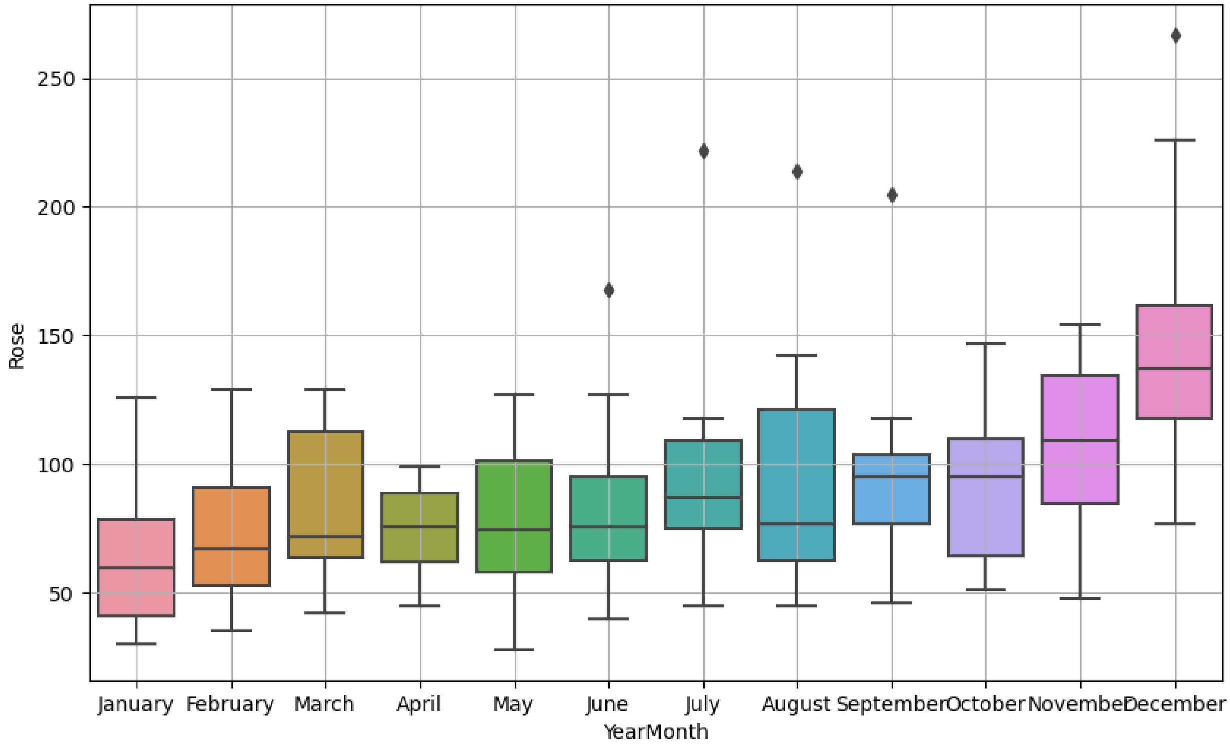
- Here, we have the statistical summary

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

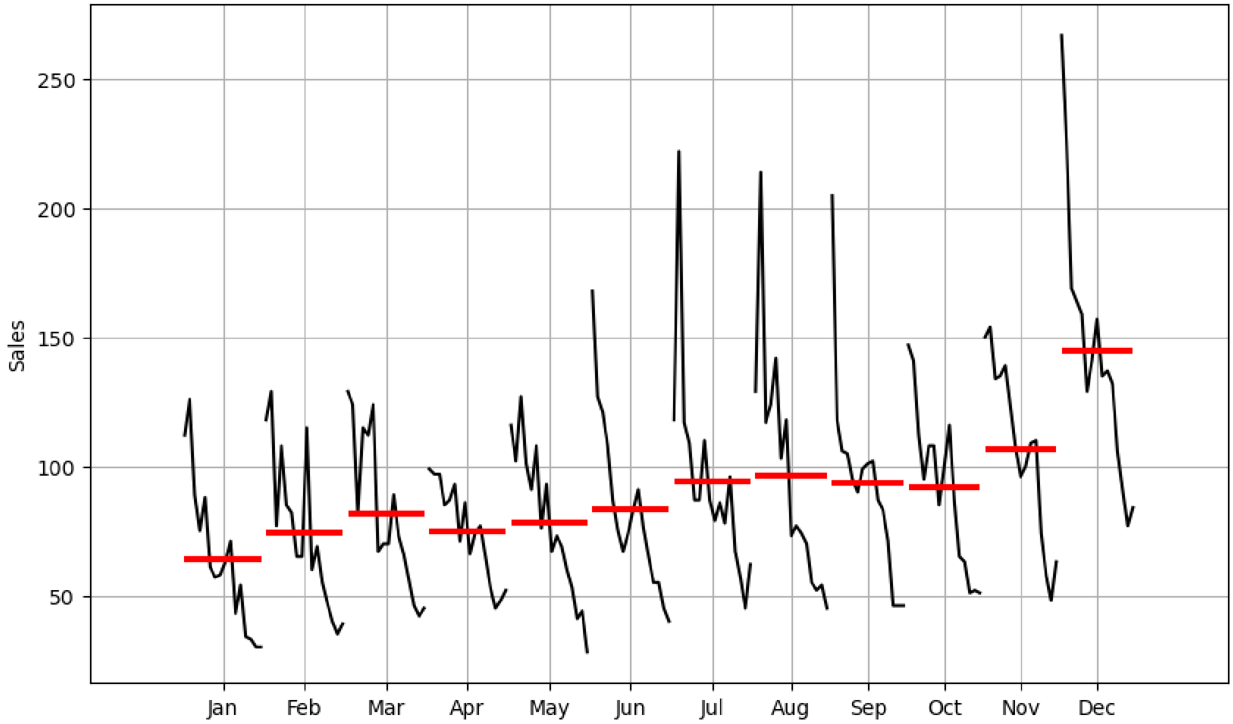
- The following is the yearly boxplot of the data



- The following is the monthly boxplot of the data



- Plotting a time series monthplot to understand the spread of accidents across different years and within different months across years

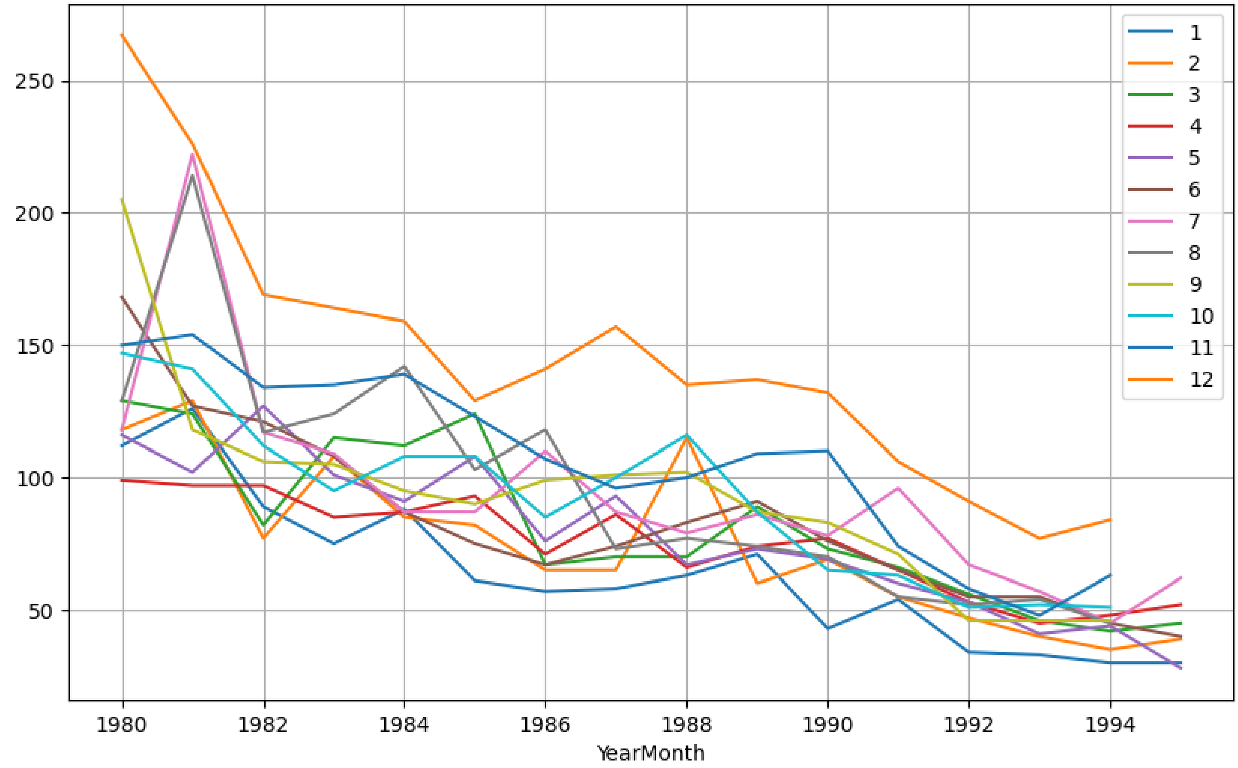


- To Plot a graph of monthly RetailSales across years, first we make a pivot table of the data

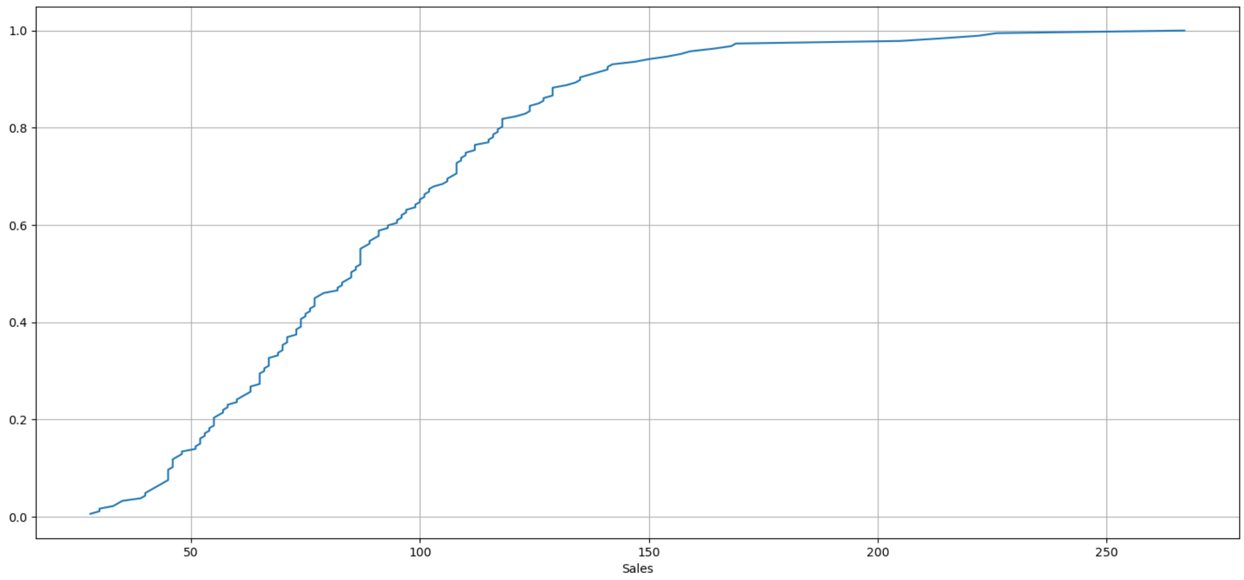
YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
[illegible]	[illegible]	[illegible]2.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.0	45.0	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

tput; double click to hide

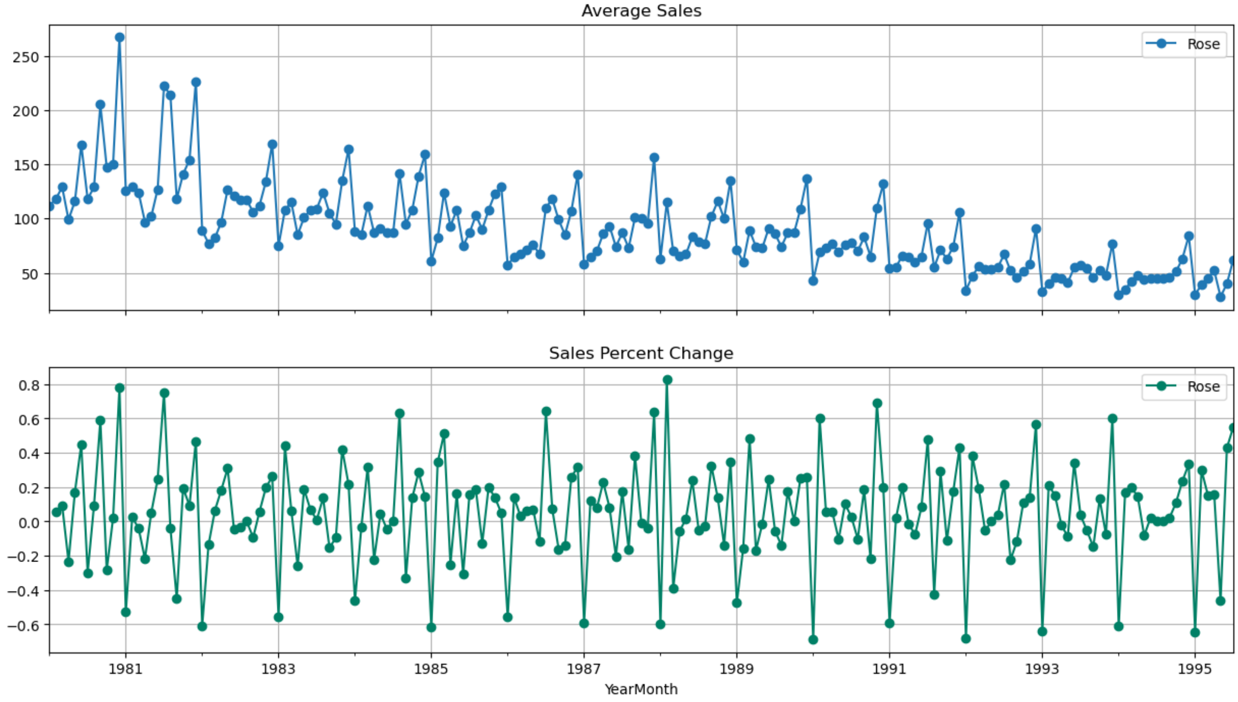
Using the above table we have plotted a line graph to understand the data better.



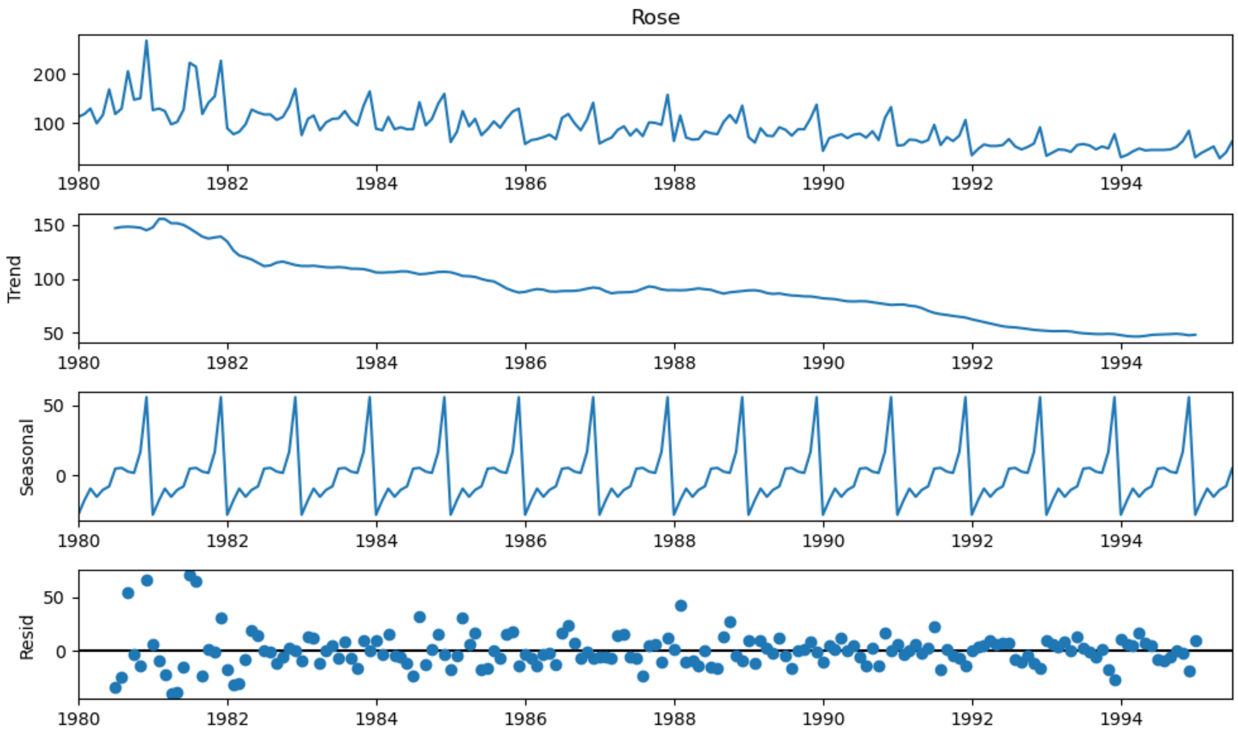
- Now we Empirical Cumulative Distribution



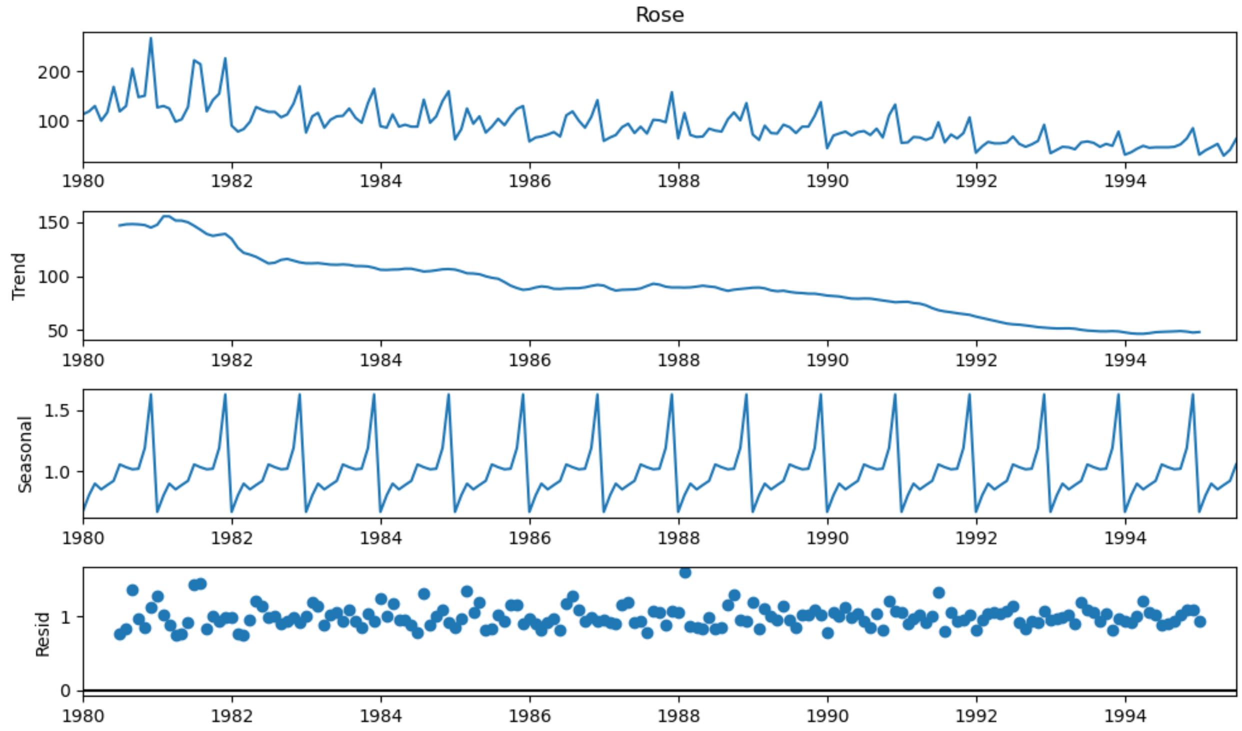
- Plotting the average RetailSales per month and the month on month percentage change of RetailSales



- In the next step, we use additive decomposition to decompose the Time Series and plot the different components



- Next, we use multiplicative decomposition to decompose the Time Series and plot the different components



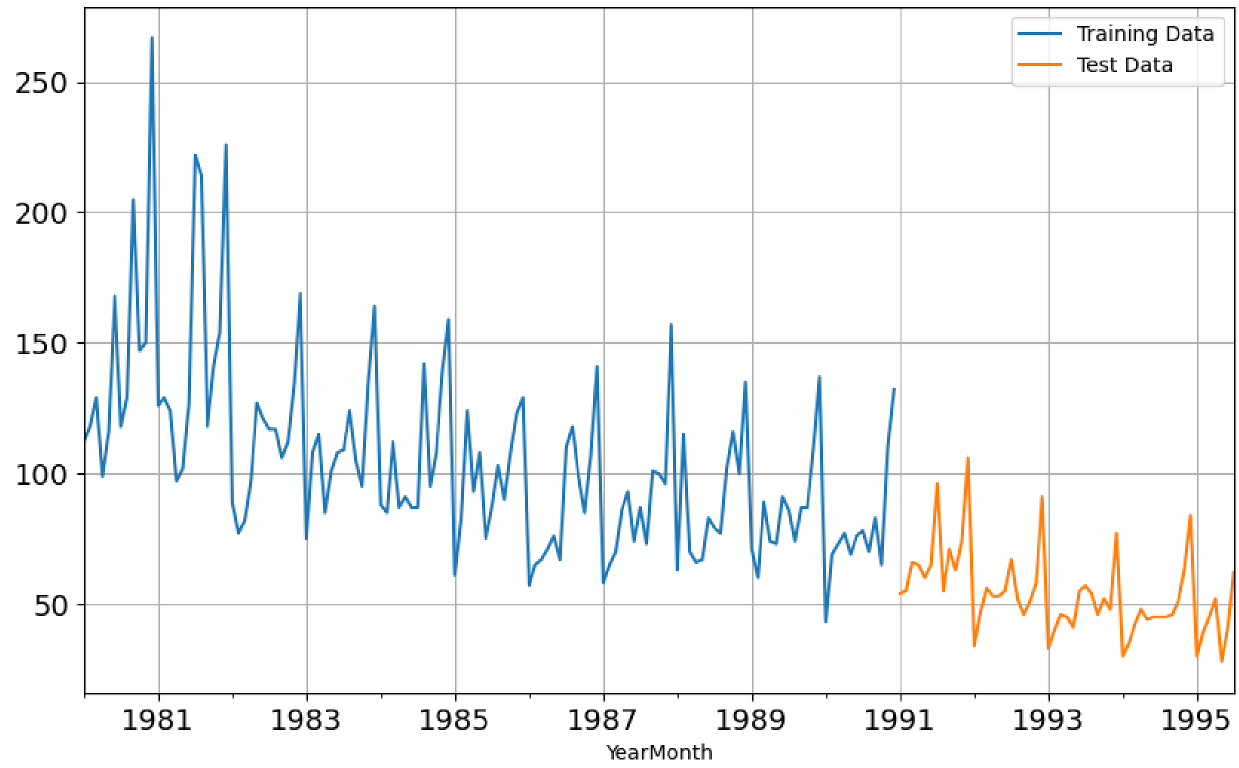
3. Split the data into training and testing. The test data should start in 1991.

Here, for training we use the data till 1990 and for testing the data post 1991 has been used.

The shape of the training data is - (132, 1)

The shape of the testing data is - (55, 1)

The following is the visualization for the train and test data split



4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

For this problem, we use multiple libraries like mean_squared_error from sklearn.metrics, ExponentialSmoothing, SimpleExpSmoothing, Holt from statsmodels.tsa.api and statsmodels.tools.eval_measures.

a. **Simple Exponential Smoothing**

The estimated parameters for SES is as follows

{'smoothing_level': 0.09874984903268463,

'smoothing_trend': nan,

'smoothing_seasonal': nan,

'damping_trend': nan,

'initial_level': 134.3869689447322,

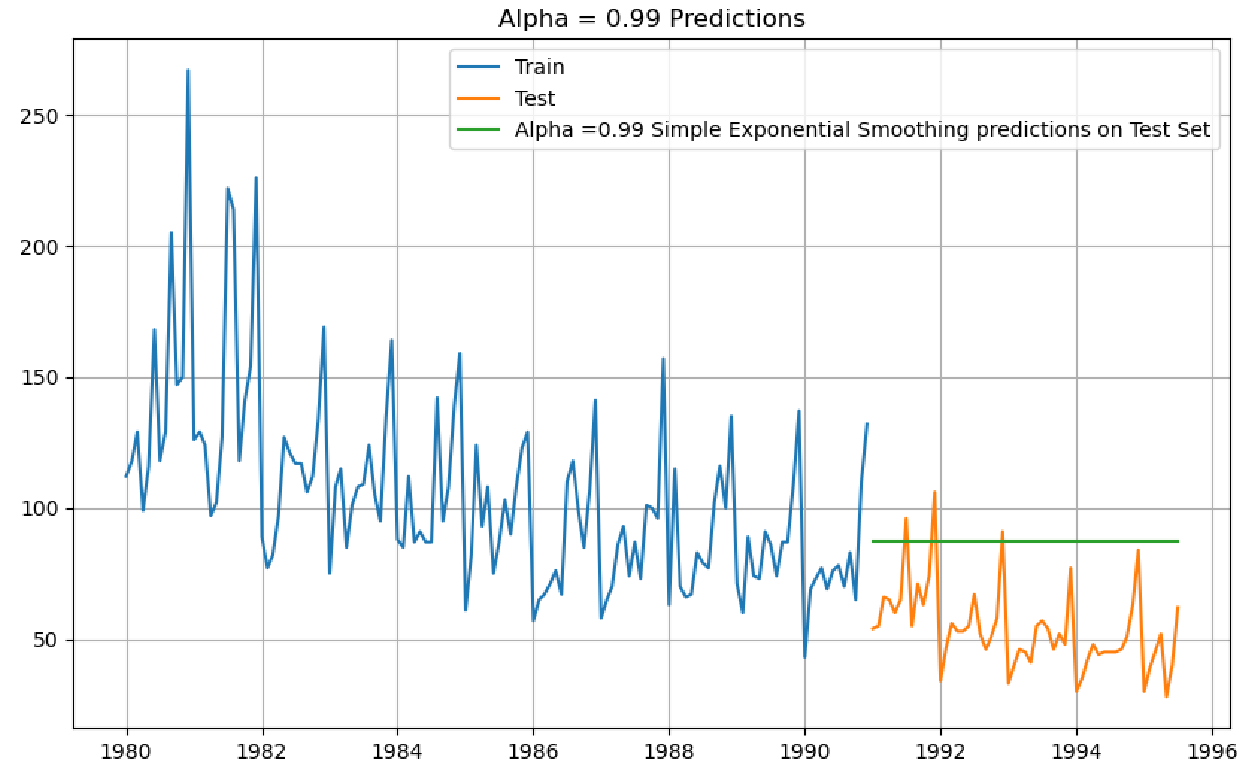
'initial_trend': nan,

'initial_seasons': array([], dtype=float64),

'use_boxcox': False,

'lamda': None,
'remove_bias': False}

Post this, we plot the following which is the simple exponential smoothing predictions on test data

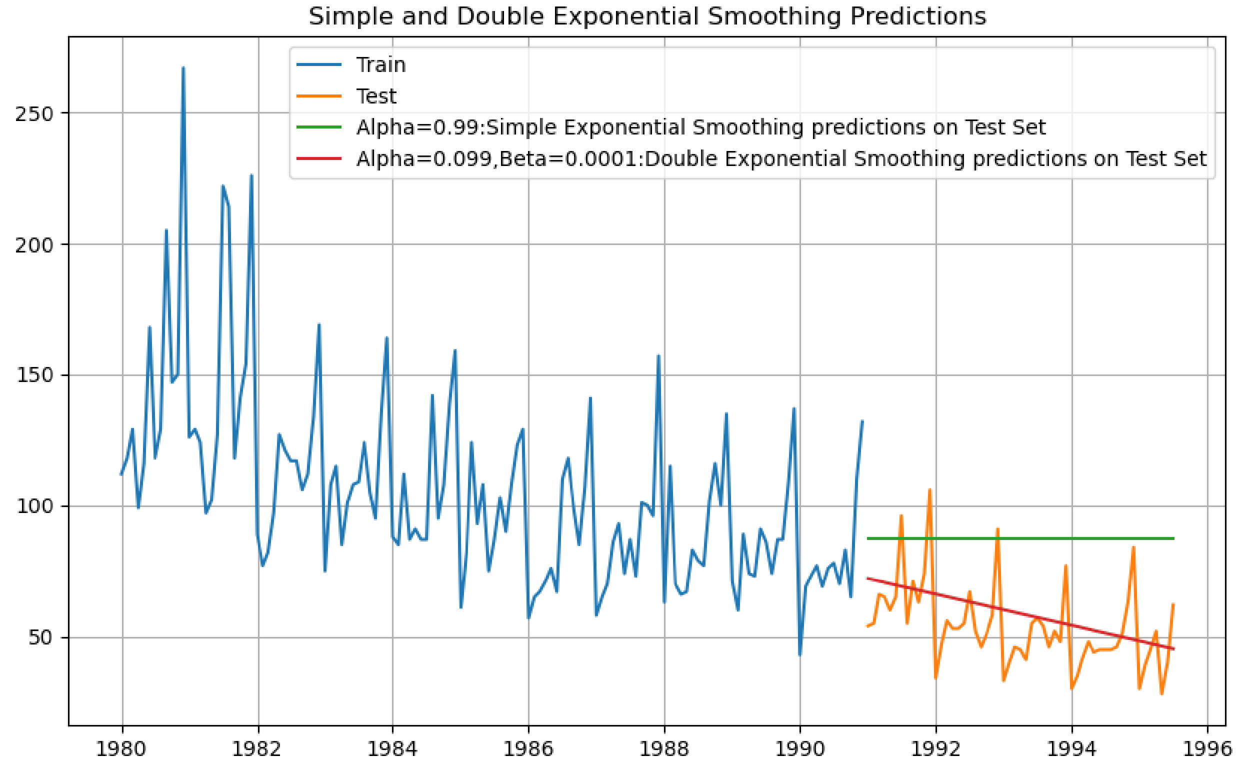


The RMSE for this model is 2750.8715511814835

b. Double Exponential Smoothing

The estimated parameters for DES is as follow {'smoothing_level': 1.4901161193847656e-08, 'smoothing_trend': 0.0, 'smoothing_seasonal': nan, 'damping_trend': nan, 'initial_level': 137.8155475230004, 'initial_trend': -0.4943783477354779, 'initial_seasons': array([], dtype=float64), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}

Post this we will plot a graph to visualize the double exponential smoothing prediction on the test data



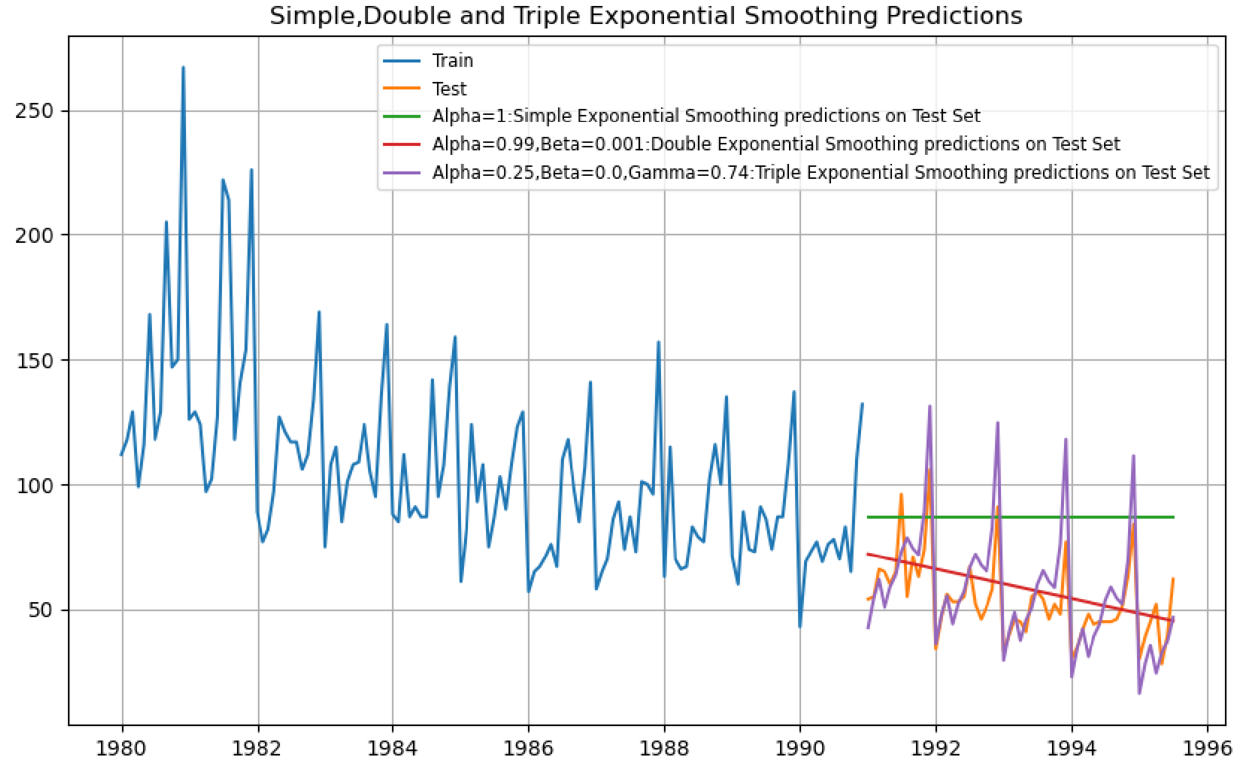
The RMSE that we have got for this prediction is 15.275715668630216

c. Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors

These are the parameters obtained

{'smoothing_level': 0.1112722708441868, 'smoothing_trend': 0.012360804011805133, 'smoothing_seasonal': 0.4607176722812185, 'damping_trend': nan, 'initial_level': 2356.5780457745022, 'initial_trend': -0.10071063842436252, 'initial_seasons': array([-636.23320656, -722.98321108, -398.64408844, -473.43046147, -808.42474493, -815.34992566, -384.23070718, 72.99480164, -237.44227544, 272.32602348, 1541.37739846, 2590.07688887]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}

The following is the triple exponential smoothing prediction on the test data



The RMSE for this model is 2609.434669404991

Now, we will build different models and compare the accuracy metrics.

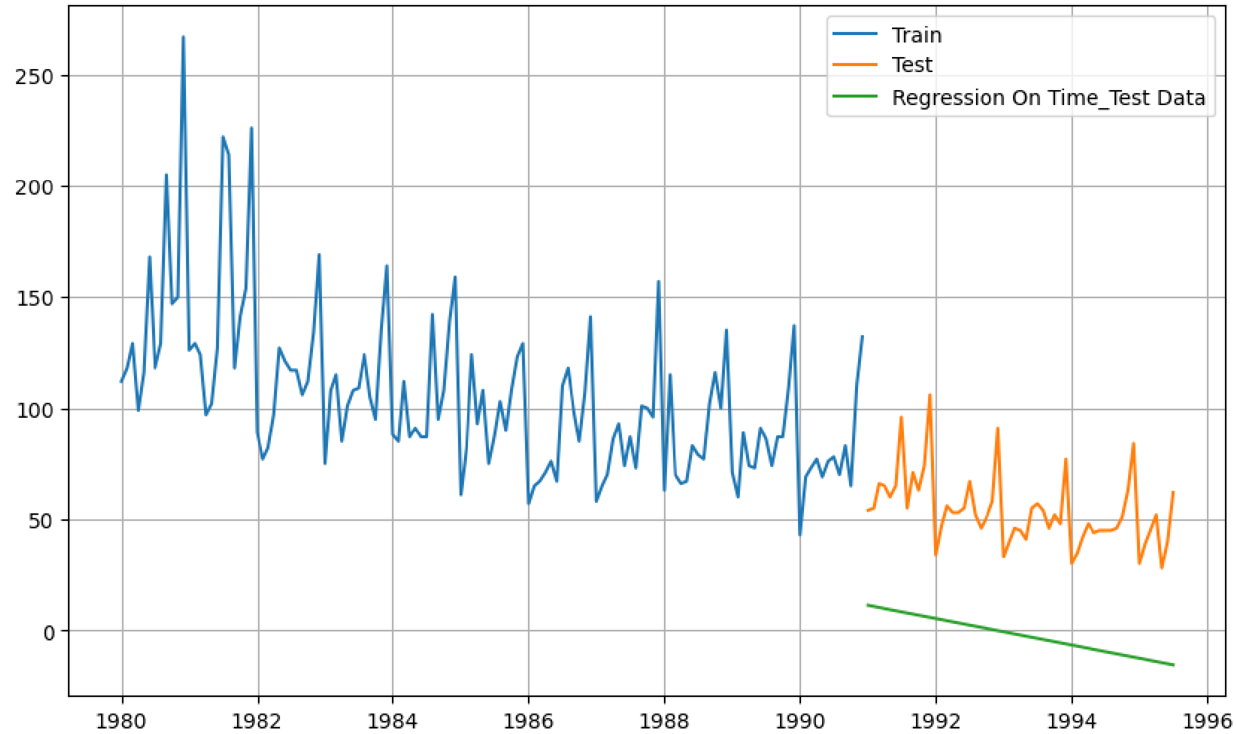
a. Linear Regression

Here, we see that we have successfully generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

YearMonth	Rose	time2
1980-01-01	112.0	1
1980-02-01	118.0	2
1980-03-01	129.0	3
1980-04-01	99.0	4
1980-05-01	116.0	5

Now that our training and test data has been modified, we go ahead and use LinearRegression to build the model on the training data and test the model on the test data.

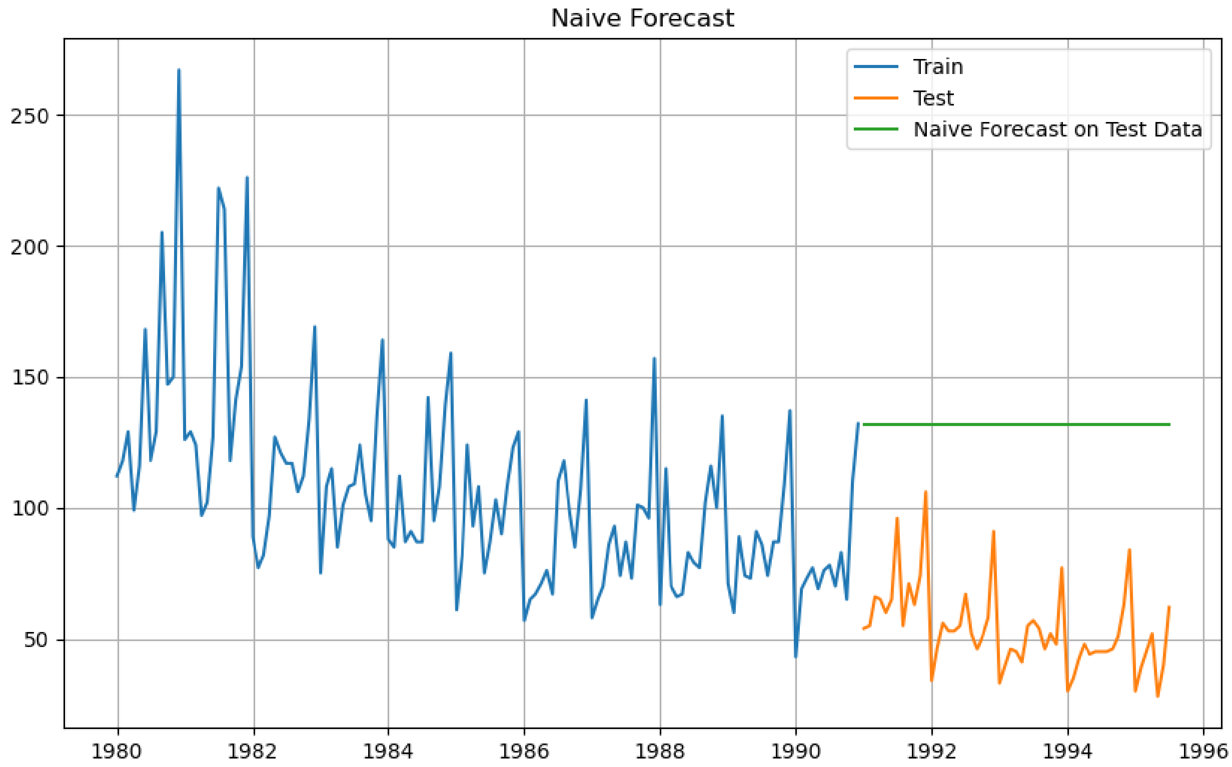
After we fit the data, the following is the output of the prediction



The RMSE for this model is RMSE is 57.773

b. Naive Approach

After fitting the data, the following is the output

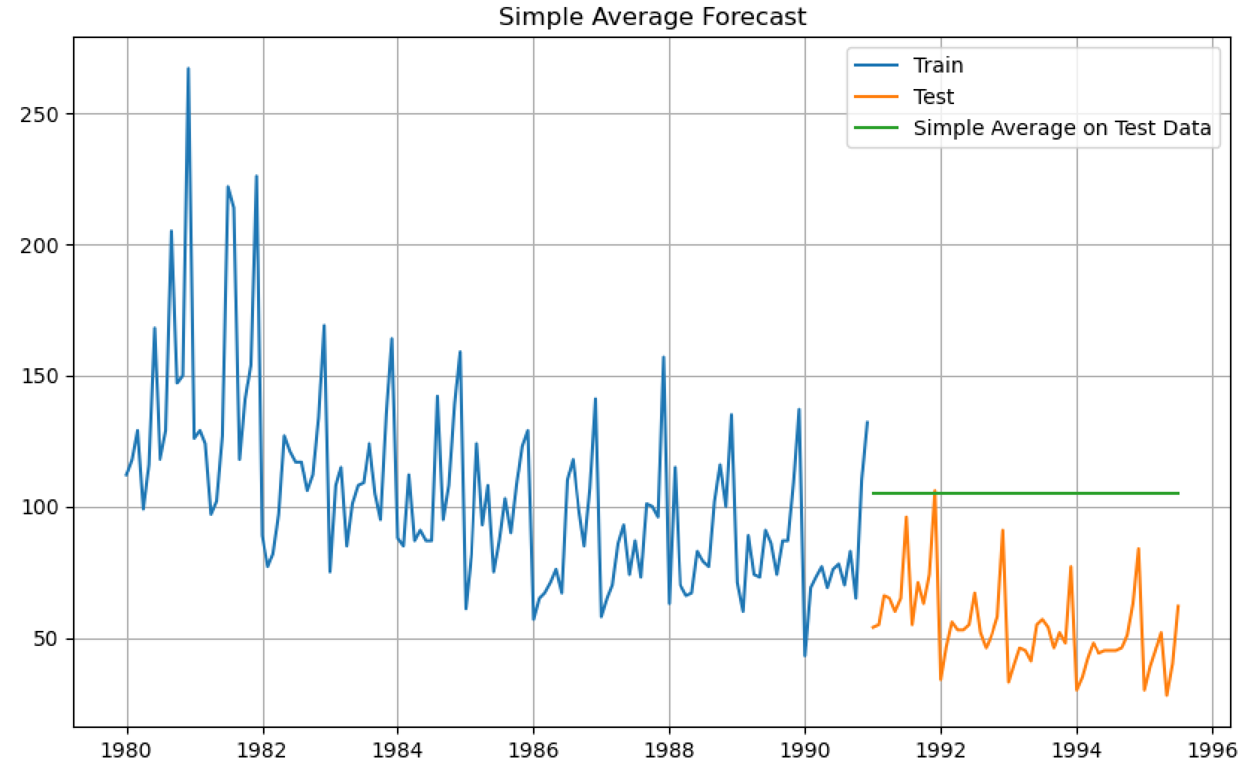


The RMSE for this approach is 79.739

c. Simple Average

This model creates mean forecast to predict the future data

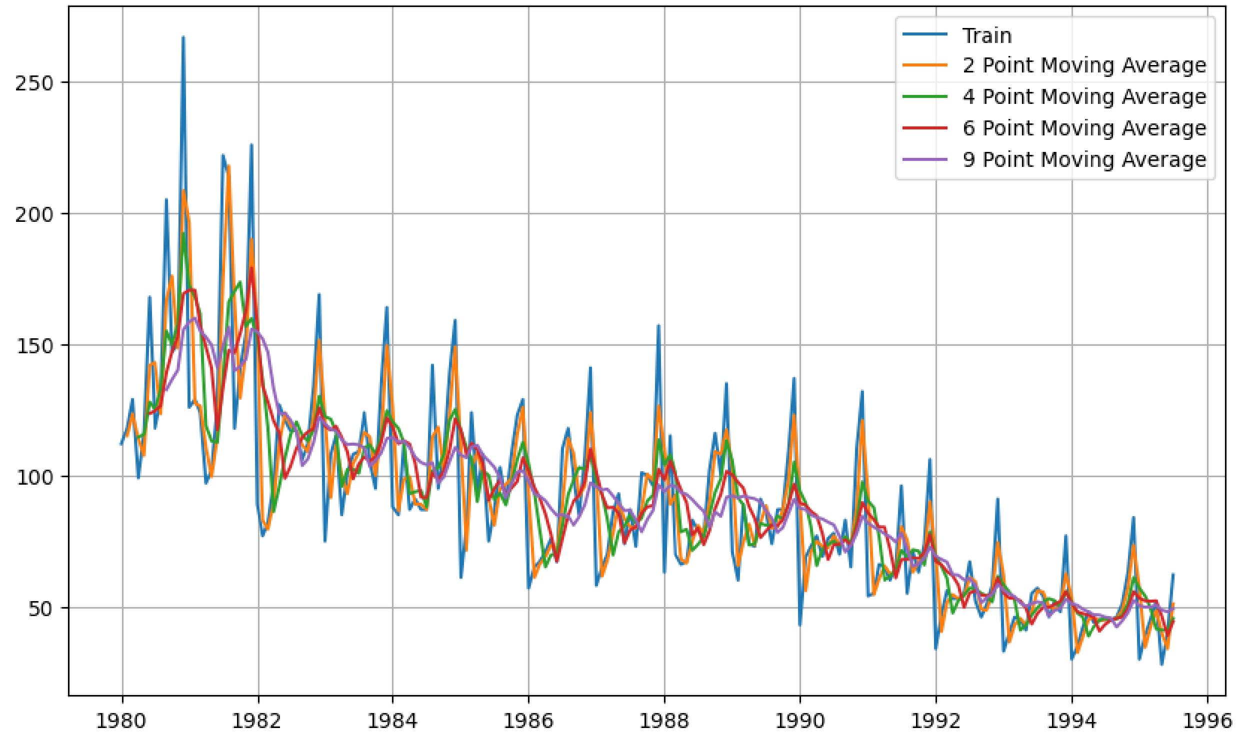
The following is the graph plotted to show the prediction on the test data



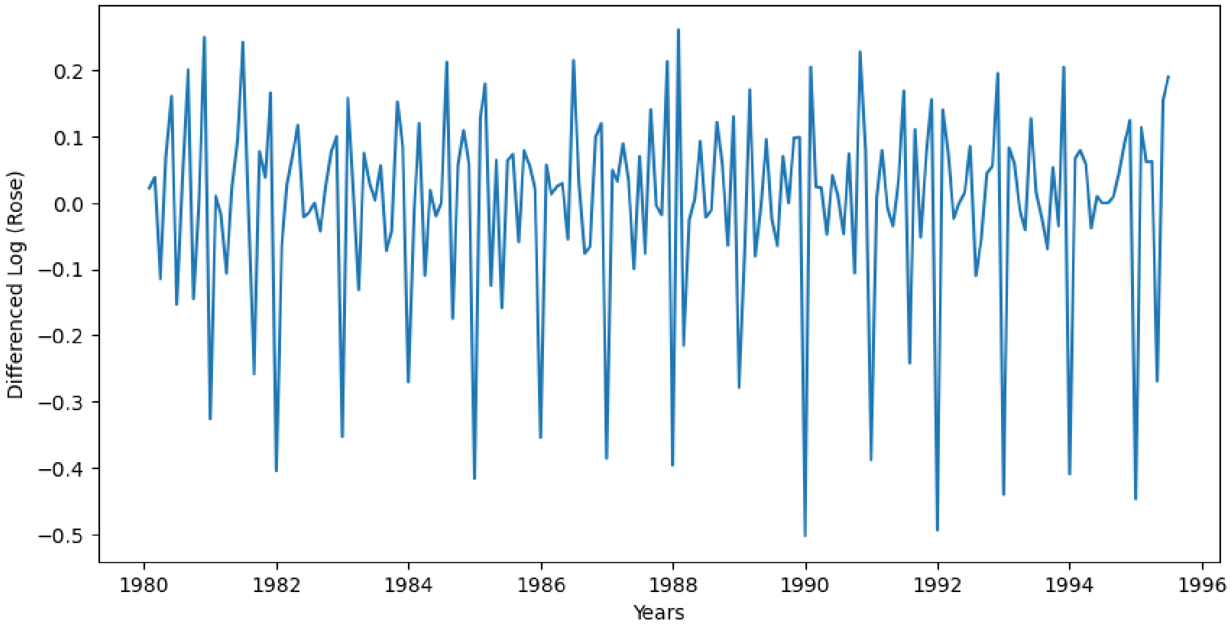
The RMSE value for this is 53.481

d. Moving Average(MA)

This is a statistical calculation that smooths out data points by averaging values within a sliding window, providing a trend or pattern representation while reducing noise.


Train
2 Point Moving Average
4 Point Moving Average
6 Point Moving Average
9 Point Moving Average
250
200
150
100
50
1980
1982
1984
1986
1988
1990
1992
1994
1996

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.



6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

a. ARIMA

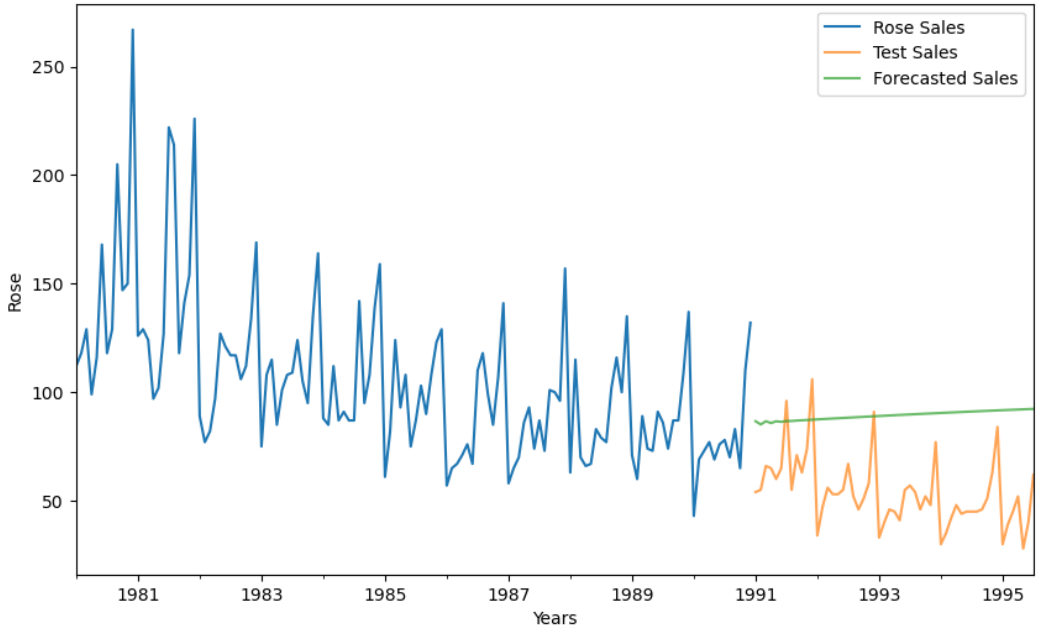
Building ARIMA model with best parameters p,d,q

	param	AIC
2	(0, 1, 2)	-186.803249
5	(1, 1, 2)	-186.195065
4	(1, 1, 1)	-186.101119
7	(2, 1, 1)	-184.829175
8	(2, 1, 2)	-184.320511

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(3, 0, 3)   Log Likelihood                  98.381
Date:                Sat, 30 Sep 2023   AIC                           -180.762
Time:                        23:02:25   BIC                           -157.700
Sample:                    01-01-1980   HQIC                          -171.391
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.0114      0.105     19.178      0.000       1.806       2.217
ar.L1         -0.1874      4.111     -0.046      0.964      -8.244       7.869
ar.L2          0.7941      1.976      0.402      0.688      -3.079       4.667
ar.L3          0.3686      2.104      0.175      0.861      -3.754       4.492
ma.L1          0.4538      4.137      0.110      0.913      -7.655       8.562
ma.L2         -0.7508      0.866     -0.867      0.386      -2.448       0.946
ma.L3         -0.4329      2.665     -0.162      0.871      -5.655       4.790
sigma2         0.0130      0.002      7.667      0.000       0.010       0.016
==============================================================================
```

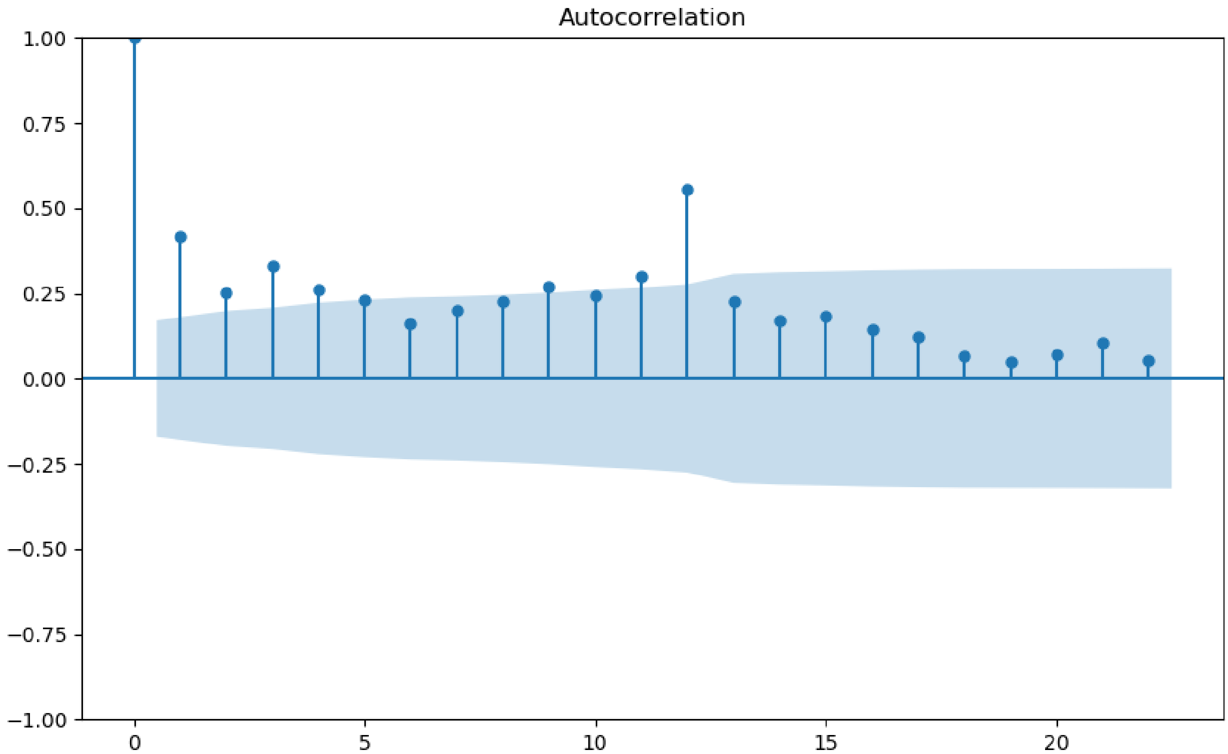
The RMSE for the above model is 531.979

The following is the plot for predicted sales



b. SARIMA Model

The start with plotting the ACF plot



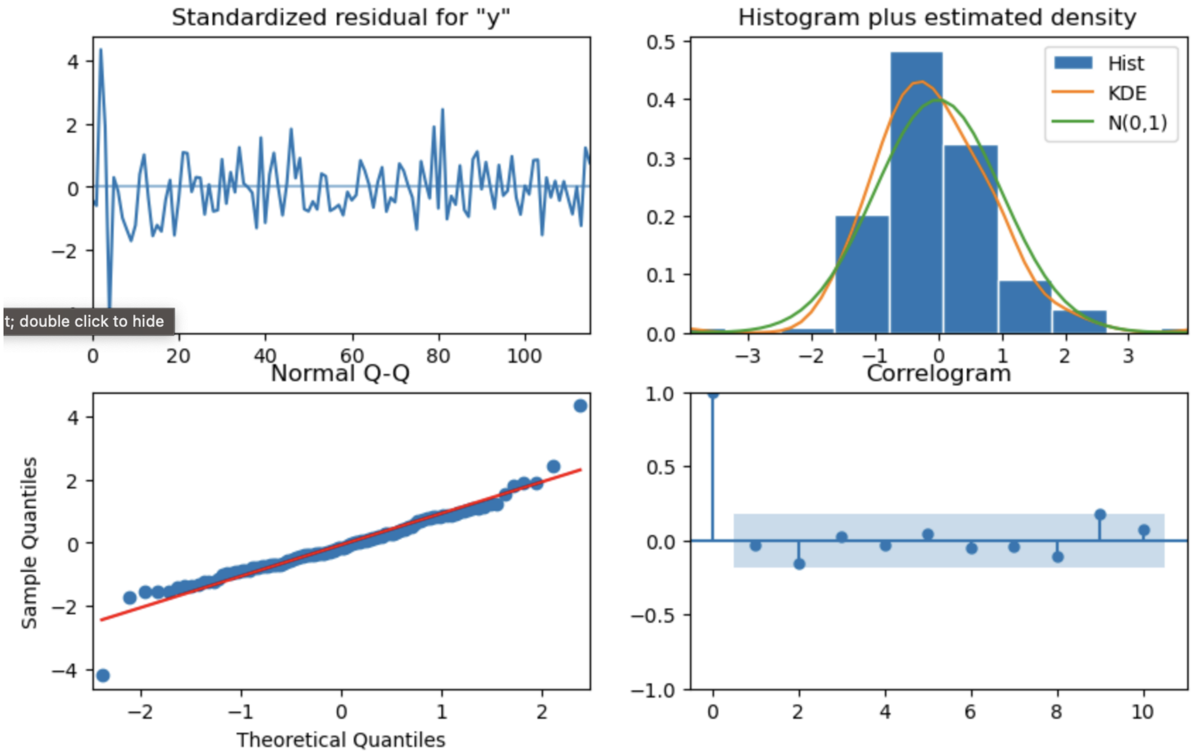
The following are some parameter combinations for Model

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1041.655818
26	(0, 1, 2)	(2, 0, 2, 6)	1043.600261
80	(2, 1, 2)	(2, 0, 2, 6)	1045.229978
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778470

The following is the summary for SARIMA

```
CONVERGENCE: NORM_OF_PROJECTED_GRADIENT_<=_PGTOL
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -514.800
Date:                          Sat, 30 Sep 2023   AIC                           1043.600
Time:                                  23:07:16   BIC                           1062.875
Sample:                                       0   HQIC                          1051.425
                                          - 132
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7884    658.323     -0.001      0.999   -1291.077    1289.501
ma.L2         -0.2116    139.298     -0.002      0.999    -273.230     272.807
ar.S.L6       -0.0727      0.037     -1.991      0.046      -0.144      -0.001
ar.S.L12       0.8368      0.042     19.880      0.000       0.754       0.919
ma.S.L6        0.2239    658.317      0.000      1.000   -1290.054    1290.502
ma.S.L12      -0.7762    510.918     -0.002      0.999   -1002.158    1000.605
sigma2       347.5179      2.606    133.371      0.000     342.411     352.625
===================================================================================
Ljung-Box (L1) (Q):                   0.14   Jarque-Bera (JB):                90.77
Prob(Q):                              0.71   Prob(JB):                         0.00
Heteroskedasticity (H):               0.42   Skew:                             0.37
Prob(H) (two-sided):                  0.01   Kurtosis:                         7.27
===================================================================================
```

After which we have plotted the diagnostic plots



The RMSE value for this model is 27.393732836163334

7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

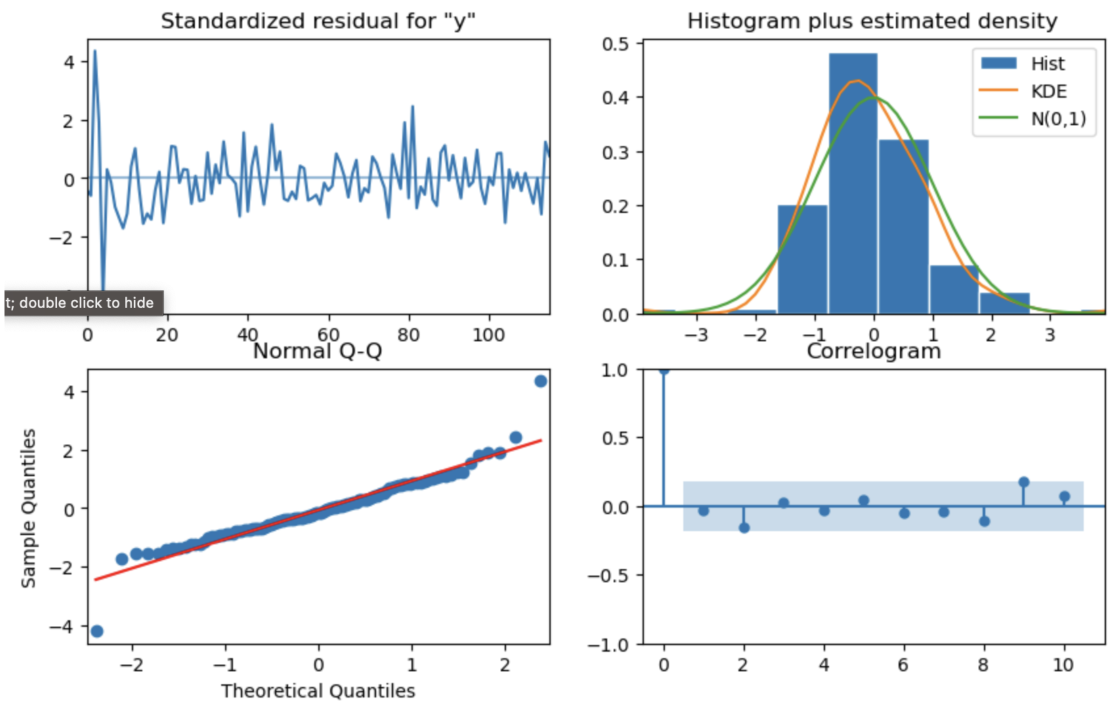
	RMSE	Test RMSE
SARIMA(0,1,2)(2,0,2,6)	27.393733	NaN
Alpha=0.99,SES	NaN	1338.004623
Alpha=1,Beta=0.0189:DES	NaN	3949.931298
Alpha=0.25,Beta=0.0,Gamma=0.74:TES	NaN	378.944325
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	NaN	403.125867
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	NaN	403.125867
RegressionOnTime	NaN	57.772711
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	NaN	403.125867
NaiveModel	NaN	3864.279352
SimpleAverageModel	NaN	1275.081804
SimpleAverageModel	NaN	1275.081804
Best ARIMA Model : ARIMA(3,0,3)	27.393733	NaN
SARIMA(0,1,2)(2,0,2,6)	27.393733	NaN

The above is a table with all the models built along with their corresponding parameters and their respective values on the test data.

8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

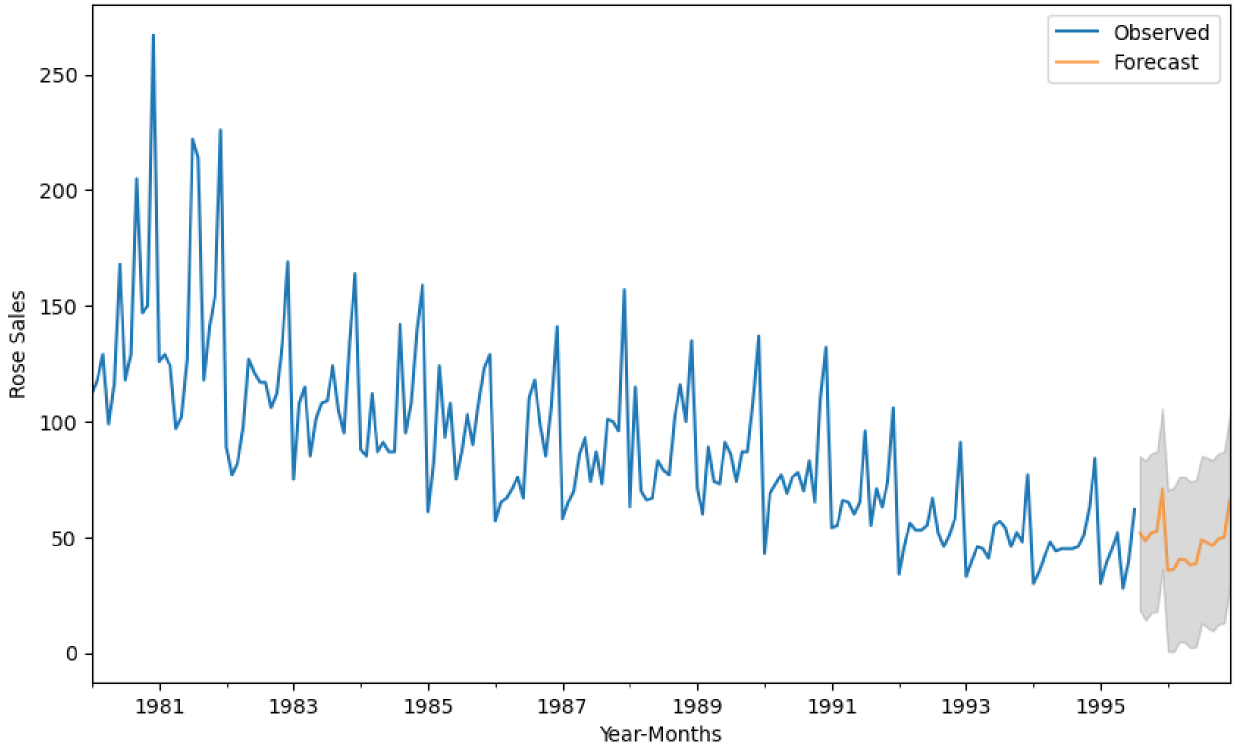
```
CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH
                                     SARIMAX Results
=========================================================================================
Dep. Variable:                           Rose   No. Observations:                  187
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -734.167
Date:                        Sat, 30 Sep 2023   AIC                           1482.334
Time:                                23:27:03   BIC                           1504.326
Sample:                            01-01-1980   HQIC                          1491.257
                                 - 07-01-1995
Covariance Type:                          opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7295      0.071    -10.344      0.000      -0.868      -0.591
ma.L2         -0.1900      0.066     -2.884      0.004      -0.319      -0.061
ar.S.L6       -0.0495      0.029     -1.684      0.092      -0.107       0.008
ar.S.L12       0.8766      0.030     29.399      0.000       0.818       0.935
ma.S.L6        0.2238      0.236      0.947      0.344      -0.239       0.687
ma.S.L12      -0.8218      0.200     -4.110      0.000      -1.214      -0.430
sigma2       269.9811     69.351      3.893      0.000     134.056     405.906
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):               297.25
Prob(Q):                              0.66   Prob(JB):                         0.00
Heteroskedasticity (H):               0.18   Skew:                             0.45
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.40
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```



The RMSE value is 28.050941294587098

The final plot of the prediction is on the entire data is



9. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

Steps involved in model building -

Data Collection: Gathering historical time series data relevant to the problem we want to forecast. Ensuring the data is complete, accurate, and covers a sufficient time period.
Data Preprocessing:

- Time Series Decomposition: Decomposing the time series into its components (trend, seasonality, noise) using techniques like additive or multiplicative decomposition.
- Handling Missing Data: Addressing missing data points through interpolation or other methods.

- Outlier Detection: Identifying and handling outliers that can distort the forecasting model.

Exploratory Data Analysis (EDA): Visualizing and analyzing the time series data to understand its characteristics, trends, and seasonality. EDA helps in selecting appropriate forecasting methods.

Data Transformation:

- Logarithmic Transformation: Applying a logarithm to stabilize variance if necessary.
-

Model Selection:

- Choosing an appropriate forecasting model based on the data and its characteristics. Common models include ARIMA, Exponential Smoothing, Prophet, or machine learning models like LSTM for deep learning.

Training and Validation:

- Splitting the data into training and validation sets. The validation set helps evaluate model performance.
- Training the selected model on the training data, fine-tuning hyperparameters as needed.

Model Evaluation:

- Using appropriate metrics (e.g., Mean Absolute Error, Root Mean Squared Error, Mean Absolute Percentage Error) to evaluate the model's accuracy on the validation set.
- Adjusting the model if its performance is unsatisfactory.

Forecasting:

- Using the trained model to generate future forecasts. The forecast horizon can vary depending on your application.

Reporting and Visualization:

- Communicate the forecast results through reports, charts, and dashboards to stakeholders.

Given the burgeoning demand in the market, ABC Estate Wines should place a strong emphasis on expanding the production of sparkling wine. Simultaneously, it's crucial to reallocate resources towards the promotion and enhancement of our Rose wine offerings, as recent sales have exhibited a decline.

Thank you!